Instr Sci
DOI 10.1007/s11251-012-9236-3

The effects of skill diversity on commenting and revisions

Melissa M. Patchan · Brandi Hawk · Christopher A. Stevens · Christian D. Schunn

Received: 30 April 2010 / Accepted: 17 April 2012

Abstract The use of peer assessment to evaluate students' writing is one recommended method that makes writing assignments possible in large content classes (i.e., more than 75 students). However, many instructors and students worry about whether students of all ability levels are capable of helping their peers. We examine how ability pairing (e.g., high-ability student with high-ability student versus high-ability student with low-ability student) changes key characteristics of feedback to determine which pairings are likely to benefit students most. A web-based reciprocal peer-review system was used to facilitate the peer review of students' writing of two papers. Over 1,100 comments given to writers from their peers were coded for several relevant categories: type of feedback, type of criticism, focus of problem, focus of solution, and implementation. Overall, creating peer-review groups such that students receive feedback from someone of a dissimilar ability appeared to be most beneficial. High-ability writers received similar kinds of feedback from high-ability versus low-ability peers. By contrast, the low-ability writers received more comments that identified problems focusing on substance issues from high-ability reviewers. In addition, the low-ability writers implemented a higher percentage of the comments from the high-ability reviewers.

Keywords Peer assessment · Grouping students · Written composition · Skill diversity · Feedback · Revisions · Peer-review

M. M. Patchan (✉) · B. Hawk · C. D. Schunn
University of Pittsburgh, Pittsburgh, PA, USA
e-mail: melissa.patchan@gmail.com

C. A. Stevens
Pennsylvania State University, University Park, PA, USA

Published online: 04 May 2012

Introduction

In higher education, students seldom have good opportunities to learn how to write because their large content classes (i.e., more than 75 students) make it very difficult for instructors to provide adequate feedback on writing (Arum and Roksa 2011; Bok 2006). One possible solution that has been gaining popularity is the use of peer assessment to evaluate students' writing in both formative and summative ways. Peer assessment can benefit students in multiple ways. First, students experience more writing assignments, which allows them to practice the skill of writing. Second, students can receive a significant amount of feedback from peers, which allows them to revise drafts more successfully (Cho and Schunn 2007). Third, students learn how to evaluate writing by providing feedback to their peers, which can improve the students' ability to evaluate their own writing (Wooley et al. 2008).

Despite these benefits, instructors are reluctant to use peer assessment. Many instructors have told us that they are unsure whether all students are capable of helping their peers. That is, instructors question whether low-ability students have something to offer high-ability students, and whether low-ability students helping other low-ability students would be too much like the blind leading the blind. Therefore, the current study addresses two research questions. The first research question focuses on a pragmatically important issue—that is, how are the benefits of peer-review moderated by the ability of the students (as writers and peers as reviewers)? The second research question builds on peer-assisted learning theory—that is, how similar does peer assessment in writing work in comparison to other types of peer-assisted learning (e.g., peer tutoring and cooperative learning)?

Rigorous empirical work on peer assessment is limited, and it tends to focus on the overall effectiveness or reliability and validity (for reviews, see Topping 1998 and Topping and Ehly 1998). One prior study (Lin et al. 2001) examined one dimension of student ability, thinking style, and found that review groups consisting of high executive thinking style students gave better feedback and improved performance relative to review groups consisting of low executive thinking style students. However, this study confounded giving and receiving feedback—high students only gave feedback to high students and low students only gave feedback to low students. The study provides no information on the effects of the two heterogeneous reviewing cases (high ability to low and low ability to high). No other ability grouping studies in the context of peer assessment of writing could be found. Thus, interactions of student ability on peer review specifically in writing have not been examined in prior work.

In a typical setting, students are randomly assigned to the peer assessment groups. Gouli et al. (2008) described several web-based environments that support peer assessment, but only their environment (PECASSE) offered alternative strategies to assign students to peer assessment groups, such as randomly by the system, manually by the instructor, or systematically by the system based on learners' characteristics—grouping students with similar characteristics or dissimilar characteristics. While a mechanism was offered to automatically create homogeneous or heterogeneous groups (as was done experimentally in Lin et al. 2001), the authors did not provide any empirical evidence or advice regarding which option was optimal. The current study examined how feedback from peers of similar ability (i.e., homogeneous groups) or from peers of differing ability (i.e., heterogeneous groups) affected the benefits of peer-review—more specifically, we examined how the features and content of peer comments differed depending on the feedback source and whether these differences affected how the students responded to the feedback. In this context, feedback is defined as the information provided to a writer with the purpose of changing performance in a particular direction (i.e., improve the quality of the document).

Given the absence of prior literature on grouping strategies specifically in writing, literature from other types of peer-assisted learning will be considered in order to develop hypotheses about how different abilities affect the benefits of peer assessment. Topping (2005) reviewed 25 years of development in the peer-assisted learning literature. He identified two types of peer-assisted learning that have been long established and rigorously researched: peer tutoring and cooperative learning. Peer tutoring and cooperative learning clearly have different goals than peer assessment, but they do all critically involve the exchange of some formative feedback, which is likely influenced by ability level (from both provider and receiver). Peer tutoring and cooperative learning also tends to be face-to-face, whereas peer assessment tends to object-centered and often online, but peer tutoring and cooperative learning does also happen in online contexts.

To map predictions, critical similarities and differences between peer assessment (our focus) and peer tutoring and cooperative learning (existing literature) must be unpacked. We consider two key dimensions: interaction and number of feedback ideas received. In the following section, comparisons between peer assessment and peer tutoring/cooperative learning on these two dimensions will be made to better understand which structural features of peer assessment are similar to peer tutoring and cooperative learning.

Comparisons between peer assessment, peer tutoring, and cooperative learning

One salient difference between peer assessment and peer tutoring/cooperative learning is the type of interaction between the students: bidirectional versus one-way interaction (see Table 1). Within peer assessment, the interaction typically is one-way; especially in the cases where web environments are used to facilitate the process, a reviewer submits feedback by a certain deadline and the writer receives the feedback after the deadline has passed without any additional interaction necessary. By contrast, in peer tutoring and cooperative learning, pairs or groups of students typically work bidirectionally either face-to-face or via chat programs. Through bidirectional interaction, communication barriers between feedback provider and feedback receiver can be reduced. The greater the initial differences between provider and receiver (e.g., language, culture, ability), the more that this communication issue is likely to be important.

Another way to compare peer assessment, peer tutoring, and cooperative learning is to consider how much feedback/instruction is provided and received (see Table 1). Based on these dimensions, peer assessment seems to be more similar to cooperative learning. Frequently, peer assessment involves providing feedback to and receiving feedback from multiple peers who notice many different problems and provide many different solutions. This multiplicity provides another opportunity for peer assessment to have advantages over

Table 1 Characteristics of the peer assessment, peer tutoring, and cooperative learning literature

	Peer assessment	Peer tutoring	Cooperative learning
Type of interaction	One-way	Bidirectional	Bidirectional
Amount of feedback	Multiple sources	Single source	Multiple sources
General findings[a]			
For low-ability students	Equally	High-ability peers	High-ability peers
For high-ability students	Equally	Low-ability peers	Equally

[a] General findings regarding which peer group would likely provide the best feedback for low-ability and high-ability students

instructor assessment. Several researchers have found that students who received feedback from multiple peers improved just as much as those who received feedback from a single instructor (for a review, Topping and Ehly 1998), and in some cases, students improved more from peer feedback (Cho and Schunn 2007). Similarly to peer assessment, students in cooperative learning situations find themselves working with several students who each notice different things. During group work, students may need to provide explanations to several peers who have less understanding of the topic to be learned. In return, these students may receive explanations from several peers to help themselves gain a better understanding. Unlike peer assessment and cooperative learning, in peer tutoring, students (i.e., the tutee) typically interact over an extended session with only one other student (i.e., the tutor). In these sessions, the tutor might provide more depth on particular issues, but is likely to notice fewer issues. As a result, peer tutoring may provide students fewer learning opportunities from receiving feedback than peer assessment and cooperative learning.

Effects of ability on cooperative learning and peer tutoring

In theory, there are several advantages to working with peers with the same ability versus working with peers with high or low ability (Lou et al. 1996). When students of similar abilities work together, the students are able to work at a common pace. Without the pressure of trying to keep up or waiting for others to catch up, students may be more motivated to learn. However, researchers have only speculated about this advantage and have not specifically examined whether this effect actually occurs in homogenous groups.

By contrast, working with peers with higher or lower ability may provide opportunities to hear from multiple perspectives, which may stimulate discussion among the group members. One study that examined diversity in the workplace found that more diversity led to conflict (Jehn et al. 1999). One thing to note is that not all conflict is negative—conflict in groups may lead to more discussion. In order to construct an outcome that will satisfy all group members, they must first engage in a discussion about each of the members' ideas. This reflection is likely to lead to a change in understanding (De Lisi and Golbeck 1999; Moshman and Geil 1998).

Cooperative learning

While many studies have examined the effects of group composition on cooperative learning, the majority of these studies focused on elementary or secondary students. One meta-analysis revealed that there was a small effect ($d = .12$) in favor of homogeneous groups (Lou et al. 1996). Interestingly, this benefit appears to be moderated by the student's ability level. That is, the low-ability students benefited more from heterogeneous groups, the average-ability students benefited more from homogeneous groups, and the high-ability students benefited equally from both homogeneous and heterogeneous groups. The results from this meta-analysis were based on studies from multiple contexts (e.g., math, science, language arts) that included wide range of ages (e.g., first grade through postsecondary school).

Relatively few studies on cooperative learning have focused on higher education students—none of these studies involved writing. The majority of these studies in higher education settings did not find an effect of group composition (Day et al. 2005; Goethals 2002; Miller and Polito 1999; Tutty and Klein 2008; Watson and Marshall 1995). Only two studies found an overall benefit of homogeneous groups over heterogeneous groups (Cobb 1999; Goethals 2001). Similar to the studies focusing on younger students, several studies reported that the benefits differed for high-ability and low-ability students. Cobb (1999) and Day et al. (2005) found that only high-ability students benefited from homogeneous groups; the group composition did not matter for the low-ability students. Tutty and Klein (2008) also found that homogeneous groups were better for the high-ability students, but the low-ability students benefited the most from heterogeneous groups.

Overall, there is not a clear pattern, except to suggest that ability-level moderation does occur. It likely needs to be investigated in each particular context because the pattern of results does not appear to generalize across all learning situations. Thus, an investigation in the context of university students providing feedback on writing is required.

If peer assessment was more similar to cooperative learning, then students should benefit the most from feedback provided by peers of similar ability. More specifically, low-ability students were expected to benefit from feedback provided by high-ability peers, but high-ability students may benefit equally from both their high-ability and low-ability peers.

Peer tutoring

Researchers have also addressed whether peer tutoring is more successful with peers with the same ability versus peers with high or low ability. This literature focuses primarily on school-age children, who either experience same-age peer tutoring (i.e., both the tutor and tutee are the same age) or cross-age tutoring (i.e., typically the tutor is older than the tutee). For this age range in which students are making significant progress each year in many academic areas, students at the same age are more similar in ability than students at different ages. For undergraduate students, there is likely to be a wide range of abilities and age is not likely to be an accurate predictor for general skills (Arum and Roksa 2011). Instead, different measures of ability should be considered.

Another benefit is the quality of explanations received. By receiving explanations from high-ability students and acting upon those explanations, the low-ability students will likely improve their understanding (Webb et al. 1995). In a meta-analysis, Cohen et al. (1982) compared studies of same-age tutoring to cross-age tutoring that included participants from first grade through twelfth grade in mostly math or reading. The tutees experiencing cross-age tutoring performed best. Thus, these results speak in favor of students working with peers who are at a different ability level. More recent work has focused on the benefits for the tutor (Roscoe and Chi 2007), but there is not a systematic comparison by age or tutor/tutee skill across studies, and thus implications for peer review matching by skill from this more recent work on learning from tutoring is unclear.

Components of feedback and revision

In understanding how student ability might influence feedback in writing, it is important to consider the details of feedback on writing and types of revisions. A recent study investigated which feedback features (e.g., summarization, identifying problems, providing solutions, localization, explanations, scope, praise, and mitigating language) were related to implementation of the feedback in document revisions (Nelson and Schunn 2009). Three of the features were associated with an increase in implementation: summarization (i.e., the reviewer recaps the writer’s main points), localization (i.e., the reviewer identifies where in the paper the problem occurs), and solutions (i.e., the reviewer offers suggestions for how to fix a problem). Therefore, feedback that includes these features will more likely be implemented.

Another recent study examined how high-ability students and low-ability students differed in their commenting style in the context of giving feedback with specific rubrics (Patchan et al. 2009). Interestingly, when provided with specific feedback rubrics, there were very few differences between the two types of peers. However, they did not examine the interactions between reviewer ability and writer ability/document quality. For example, from a perspective of Vygotsky's (1978) zone of proximal development, low ability reviewers may not be as able to find problems in higher quality documents than high ability reviewers.

Competing predictions

Previous research has provided mixed results on how to best assign peer assessment groups. Different predictions could be made depending upon which literature was examined (see Table 1).

Cooperative learning

- If peer assessment was more similar to cooperative learning, then students should benefit the most from feedback provided by peers of similar ability. More specifically, low-ability students were expected to benefit from feedback provided by high-ability peers, but high-ability students may benefit equally from both their high-ability and low-ability peers.

Peer tutoring

- If peer assessment was more similar to peer tutoring, then students should benefit the most from feedback provided by peers of different ability. However, this prediction is very tentative given the lack of ability grouping studies on the benefits for the tutor.

Writing research

- Based on what is known from writing research, students may not benefit more from same ability or different ability peers because the commenting style between the two groups seems very similar.

We explored these predictions in a setting in which high and low writing ability students as authors received feedback from high and low writing ability students as reviewers; we teased apart the relative contributions from each reviewer through a careful analysis of comments received in each review and the revisions made on the basis of that comment content.

Method

Overview

The purpose of this study was to examine how student ability affected peer assessment in writing. Random assignment to peer assessment groups was used in a large Writing Across the Curriculum (WAC) course. After the course was completed, we determined the ability level of the students based on self-reported SAT verbal scores and looked for differences in the amount of feedback, feedback content, and implementation rates between high-ability and low-ability students. Peer assessment of writing was chosen as the context because of the relative ease of data collection and the large amount of data available for analysis. This context also provides high external validity. The data were segmented and coded for analysis in order to determine whether statistically significant relationships existed between the writing skill levels and the rates of feedback features and implementation.

Course context

With an increase in WAC programs, content courses are more likely to use peer assessment in order to add writing assignments to large lecture courses. Therefore, this study was conducted in an Introduction to Cognitive Science course at a top-tier mid-sized public research university in the US. There were two main goals for the course: to give students a general understanding of the research topics and scientific methods used in cognitive science, and to provide students an opportunity to derive a more detailed understanding of the areas that overlap with their long-term interests. In order to accomplish the second goal, two five-page papers were assigned. Students applied a key scientific finding or theory from one of the covered chapters (i.e. logic, rules, concepts, analogies, images, connections, brains, or emotions) to everyday life, and thus the students were writing papers that involved summarizing prior research (relatively little criticism or praise), and creating novel extensions of this prior research that required supporting logical arguments. This study examined the combined data set of both papers.

Participants

An extreme-groups comparison approach was used for this study. A student's ability was determined from the self-reported SAT verbal score.[1] From a total of 91 students enrolled in the course, only 78 students reported their SAT verbal score. Based on the distribution of the scores, the top 25 % were considered high-ability participants (i.e., 20 students with an SAT verbal score of 700 or higher; $M = 743$; $SD = 34$) and the bottom 25 % were considered low-ability participants (i.e., 18 students with an SAT verbal score below 600; $M = 533$; $SD = 40$). It is important to note that we use the terms high and low in a relative sense, and that all of the participants in this study are at or above national averages of writing ability given the relatively selective nature of this particular university context. In order to compare writer and reviewer pairs, the terms *writer skill* and *reviewer skill* were given to refer to the writing ability of the students when they acted as writers and the writing ability of students when they acted as reviewers, respectively.

This study utilized a high-low extreme-groups design to carefully analyze comments and implementation of comments. Extreme-groups comparisons provide greater statistical power without bias, and are particularly useful for labor-intensive studies such as those involving detailed comment coding. Extreme-groups comparisons are also a commonly used strategy in the expertise literature for identifying strategies differences. Because the students are drawn from one relatively selective university context rather than broadly

[1] In other projects investigating writing in college settings, we have found correlations over $r = .8$ between self-reported and actual SAT scores. We have also found strong correlations between SAT verbal and more direct measures of writing ability (e.g., $r = .8$ between SAT verbal and SAT writing). We use SAT verbal because students left this blank least often. As an additional validity check, in the current dataset, we found that, SAT verbal correlated very well with TA grades ($r = .48$) and peer ratings ($r = .51$). Critical to our high/low ability groupings, differences by group were 1.3 standard deviations on final draft scores as graded by the class teaching assistant (i.e., a very large effect size).

sampled across many university contexts, the issue of generalizability of effect sizes to the full range of students is moot.

Data from 38 students (61 % female) were included for analyses. Because this course was a survey course, there was a diverse population. While undergraduates at all levels were represented, the majority of the students were lower-level undergraduates (36 % freshman, 22 % sophomores, 28 % junior, 14 % senior). In addition, students came from many different majors (24 different majors were represented). A large portion of the students majored in humanities (47 %) or natural sciences (18 %).

Procedure

Data from two writing assignments were collected and analyzed. There were four steps to the assignment that repeated for each paper: (1) submit first draft, (2) review peers' papers, (3) revise draft based on peers' feedback, and (4) back-review the helpfulness of peers' comments. To facilitate the anonymous review process for both papers, students used the Scaffolded Writing and Rewriting in the Discipline (SWoRD, version 3.5) system, a web-based reciprocal peer-review system (Cho and Schunn 2007).

First, students uploaded their first drafts to SWoRD by the specified deadline. Once the deadline passed, the SWoRD system randomly distributed five papers to each student. Then each student had 1 week to review the papers he or she was assigned. The reviewing task consisted of two parts: feedback and ratings. The students were provided with a detailed rubric consisting of three dimensions (i.e., prose flow, argument logic, and insight—see Cho and Schunn 2007 for rubric details). These dimensions were the default in the SWoRD system; the role of these dimensions in our results will be addressed in the general discussion. The students were instructed to focus on the high prose issues that were described in the rubric (e.g., "Did the author just make some claims or did the author provide some supporting arguments or evidence for those claims?"); they were asked to only comment on low prose issues, such as typos and simple grammar problems, if they were so bad that it was hard to understand the text. Written feedback was in the form of end comments (i.e., comments separate from the text) rather than marginalia (i.e., comments located in the margins of the text). The numerical rating was based on a 7-point scale (1-Disastrous to 7-Excellent). The students entered all of the comments and the ratings in SWoRD before the specified deadline.

After the reviewing deadline, each student was able to access his or her reviews, which consisted of comments and ratings from five peers. The students were then instructed to use the feedback to revise their papers. After the submission of the second draft, the students also generated written back-reviews to each of the five reviewers regarding the helpfulness of the feedback. In addition, the writers rated the helpfulness of each review on a 7-point scale (1-least helpful to 7-most helpful). These back-review comments and ratings were also entered in SWoRD.

In order to provide students with a strong incentive to take the peer-review process seriously, a large portion (40 %) of their final grade depended on their performance in both writing and reviewing activities. The writing grade was based on the average of the students' ratings for the first draft and the teaching assistant's rating for the second draft. The reviewing grade was based on two factors: how reliably the students rated their peers' papers in comparison with the other peers who rated the same papers (i.e., a score calculated automatically by the SWoRD system) and how helpful their peers found their comments (i.e., the average of the ratings provided by the peers). The draft ratings provided by each peer was kept anonymous in order to prevent the helpfulness ratings from being biased by the tit-for-tat phenomenon (i.e., students who received low draft quality ratings providing low helpfulness ratings in retaliation).

Coding process

The feedback provided in the peer-reviews was coded at a detailed level to determine how the quantity and nature of feedback varied as a function of reviewer and writer ability. Because reviewers were randomly assigned, a participant often received reviews from a combination of high-ability, medium-ability, or low-ability peers. If the comments across reviews were generally the same, it would be hard to attribute the effect of a particular review on the document revision and thus attribute revision success with writer/reviewer ability pairings in this mixed feedback setting.

A random subset of papers from the current dataset was chosen to examine the relationship between different peer comments on a given paper. Overall, we found that tended to receive supplementary comments 85 %, rather than overlapping comments 9 % or contradictory comments 6 %. Thus revisions can typically be attributed to a particular review, which is critical for our current analyses.

Because we were only interested in the extreme-group comparisons, we further reduced the number of drafts examined. At the level of drafts, we sampled only drafts that were reviewed by high and/or low ability peers. This sampling procedure excluded seven drafts that were reviewed only by medium-ability students. Therefore, out of the 76 possible drafts (i.e., 38 students in the extreme groups times two drafts each), only 69 drafts were included in the analyses.

For the papers included in the analyses, not all comments were included, but rather only those comments that came from high or low ability reviewers. Thus, each paper received anywhere between 1 and 5 total reviews across high- and low-ability reviewers, with the reviews from mid-ability peers excluded. This review selection process produced a total of 165 reviews from either high or low ability reviewers to the selected 69 drafts.

The unit of analysis for this study was the review, which consisted of all the comments from two of the reviewing dimensions (i.e., prose flow and argument logic) provided by a single reviewer. A total of 330 pieces of feedback were analyzed (i.e., 165 reviews × 2 dimensions = 330 pieces of feedback). Because reviewers commented on several issues within a single dimension, each piece of feedback was subdivided into comments that were based on idea units. For example, a comment about the clarity of the main idea was separated from the comment about needing better transitions. The number of comments within a single piece of feedback (i.e., all the comments written by one reviewer about one of the dimensions) ranged from one to nine comments. Consequently, 1,138 comments were coded.

The task of coding the type of comment was divided between two research assistants (RAs). First, the RAs coded a random subset of 230 comments to establish reliability. After coding these comments separately, the RAs discussed discrepant codes to come to an agreement on the most appropriate code. For two types of comments (type of feedback and type of criticism), reliability on the initial set of comments reached a kappa >.8, so the RAs divided the remaining segments and coded them independently. For the other three types of feedback (focus of problem, focus of solution, and implementation), the kappa was at acceptable levels (above .6) but <.8. To increase the effective reliability of the coding for these dimensions, all comments were double coded. For the double-coded comments, all discrepant codes were discussed in order to select the appropriate code.

The coding scheme established by Nelson and Schunn (2009) was used to categorize the types of comments (see Table 2 in Appendix for definitions and examples). Based on this coding scheme, the feedback features of each segment were coded at multiple levels.

Type of feedback

First, all of the segments were coded as *summary*, *praise*, or *criticism* (kappa = .97; see Fig. 1). These features were especially important because Nelson and Schunn (2009) found that feedback with summary statements were more likely increase a student's understanding of the problem, which in turn would increase the likelihood of those comments being implemented. Therefore, it was important to distinguish this comment type from the others. Students were instructed to summarize the main points of the paper, so comments that did so were coded as *summary*. For example, one student stated, "Her paper describes rules and the methods we use to engrave them into our brain, the way we learn as humans, and the way other animals learn." Comments that focus on a positive feature of the paper were considered to be *praise* (e.g., "The counterpoints were also very efficient, which were emphasized with those concrete examples."). Finally, *criticism* was used when the reviewer identified something wrong with the paper or suggested a possible improvement. For example, one student commented, "The introduction runs right into the results of the study, making it confusing. The author should separate the first half of the paper by more paragraphs."

Type of criticism

The next level to be coded was the type of criticism; only the *criticism* comments were further coded as *problems*, *solutions*, or *both* (kappa = .86; see Fig. 1). Again, these features were important because Nelson and Schunn (2009) found that comments with solutions were more likely to be implemented. Therefore, we wanted to determine whether this effect differed depending on the writer's ability and/or the reviewer's ability. The

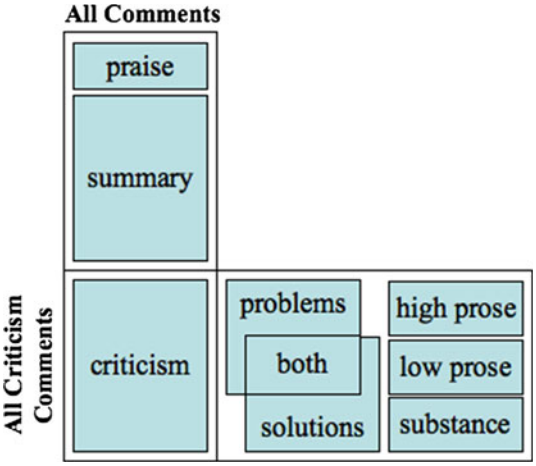


Fig. 1 Illustration of the relationship among the types of feedback coded—praise/summary/criticism, problems/solutions, and high prose/low prose/substance

comment was considered to contain a *problem* when the reviewer explicitly describes what is wrong with the paper (e.g., “There is only one problem. I’m not sure what topic you are arguing. Are you talking about logic, rules, or concepts?”). Likewise, the comment was considered to contain a *solution* if the reviewer provides an explicit suggestion for improvement (e.g., “In the first paragraph, they explained the direct quotes well to show understanding, but I think they should start off the paper with their own idea instead of using a quote.”). If the reviewer described a problem and offered a solution about the same idea, the comment was coded as *both*.

Focus of the feedback

The focus of problem and the focus of solution were considered separately (see Fig. 1): all segments that were coded as *problem* or *both* were further categorized as being focused on *high prose*, *low prose*, or *substance* (kappa = .70), and separately all of the segments coded as *solution* or *both* were coded as being focused on *high prose*, *low prose*, or *substance* (kappa = .74). This addition to the Nelson and Schunn (2009) coding scheme was necessary because we also wanted to focus on the content of the feedback rather than just the features. The low prose category was also included even though students were instructed not to include them because students had a difficult time completely ignoring all low-level issues. Therefore, we wanted to be able to show that even thought students include low prose comments, the majority of their comments focus on high-level issues. A total of 723 *problems* and/or *solutions* were coded. A *high prose* comment was one that focused on the reviewing dimensions (i.e., prose flow, argument logic, or insight). For example, one student commented on the logic of the paper: “You didn’t seem to make any counter arguments though and you might want to add those for your final paper.” If a comment focused on low-level writing issues, such as grammar or word choice, the comment was coded as *low prose* (e.g., “Also, instead of using many “him/hers” in your third paragraph, try to change the wording a little bit […]”). When comments focus on issues that only someone with content knowledge could fix, the comment was coded as *substance*. For example, one student commented, “You didn’t tell the reader about the first research done by Rosch and Mervis, what was done, how these ideas were proved. It would help the reader understand your argument if the research was there to support it.” It was possible that the problem and solution were focused differently. Consider the following comment: “I think the little red riding hood example was a little unclear, try to break the sentence down into two sentences to make it a little easier to understand.” The first part of comment described a high prose problem (i.e., the example lacked clarity), but the second part of the comment offered a low prose solution (i.e., breaking up a sentence is a minor textual repair). For five of the *problems* and/or *solutions*, the comment was off topic. These comments were disregarded.

Implementation

Finally, identified as an important dependent variable in the Nelson and Schunn (2009) coding scheme, we also coded all *criticism* comments for implementation (i.e., *implemented* versus *not-implemented*; kappa = .65). For each *criticism* comment, the RAs searched the paper to examine whether the problem had been corrected or the solution had been executed. To facilitate the search, the RAs used Microsoft Word’s compare documents function to highlight the differences between the first and second draft of the paper. As long as the writer attempted to make a revision based on the comment, that comment was coded as *implemented*. Less than 5 % (19 segments out of 503) of the segments were coded as being too vague to determine whether they had been implemented. For example, one student commented, “Flow is the area in which I think your paper needs work.” These comments were excluded from implementation coding.

While many revisions made by students involved mechanical or micro-level revisions, the RAs were able to code for macro-level revisions as well. For example, one reviewer commented, “You had a great iintro [sic.] to research to application but then your paper just stopped. You need a much better conclusion. Just restate your paper and how combining subjects would make learning smoother.” In response, the writer added a conclusion, which involved added a whole new paragraph (four sentences, 120 words).

In order to transform the raw comments into quantitative data, two steps were completed. First, the number of comments that a writer received from one reviewer for each feedback feature (e.g., *criticism*, *praise*, *summary*) was calculated. Second, for each writer, the average number of each feedback feature was calculated separately for high-ability reviewers and low-ability reviewers. In order to produce sufficiently powered analyses, the data was collapsed across the two papers. Additional analyses verified that the results were similar for each paper. For each category of feedback/implementation, outliers that were >2 standard deviations from the mean of that category were excluded.[2]

Results and discussion

Overview

The intention of this study was to understand how student ability influenced peer feedback. Specifically, we wanted to know which ability pairing (e.g., high-ability student with high-ability student versus high-ability student with low-ability student) influenced writers to implement more peer feedback (especially involving more substantial revisions), which may be a function of how much and what kind of feedback was received. Therefore, we analyzed the amount and type of feedback that reviewers of different ability levels gave writers of different ability levels, as well as the writers’ implementation rates. Because data was collected from a course that randomly assigned students to the peer assessment groups, students received feedback from high-ability reviewers and/or low-ability reviewers. As a result, it was not possible to directly assess the differences reviewer ability would have on the final quality of the paper (i.e., for a given paper, it would be unclear whether an improvement in quality was a result from the high-ability reviewers or low-ability reviewers). Instead we examine the process data, focusing on kinds of feedback thought to be most useful for quality revision of writing.

The average number of each feedback feature was analyzed using a 2 × 2 mixed-design ANOVA with writer ability (i.e., high-ability versus low-ability) as the between-subjects variable and reviewer ability (i.e., high-ability versus low-ability) as the within-subjects variable. For significant interaction effects, t-tests were performed to compare high-ability versus low-ability reviewers for high-ability writers and low-ability writers separately (see Table 3 in Appendix for means, standard deviations, and effect sizes). As it happens, the reviews did not differ in quantity by writer or reviewer ability (overall, $M = 6.9$ comments per review; $SD = 1.4$). There were no significant main effects or interaction. The lack of an effect on number of comments is convenient for the in-depth analyses of the comments,

[2] Overall, <5% of the data was removed.

as it is not necessary to distinguish between proportion of a review that contains a particular comment type and the raw number of comments of each type. Unless indicated otherwise, all reported analyses were based on raw numbers.

The effects of writer and reviewer ability did influence the type of feedback produced. Three patterns of data were commonly found in many of the analyses. First, there was a main effect of writer ability—low-ability writers received more feedback and implemented more feedback than high-ability writers. Second, there was a main effect of reviewer ability—high-ability reviewers provided more feedback and their feedback was more likely to be implemented than low-ability reviewers. However, both main effects were driven by one cell in the interaction between writer ability and reviewer ability—there was no difference for high-ability students, but low-ability students received more feedback and implemented more feedback from high-ability students than low-ability students. Because the interaction was what drove the results, the remaining sections will focus only on the interactions (see Table 3 in Appendix for summary of all statistics performed).

Type of feedback

First, we were interested in the differences in the relative amounts of major types of feedback (e.g., *summary, praise, criticism*). Summary comments were offered the least (overall, $M = .8$ comments per review; $SD = .4$). There were no significant main effects or interaction for *summary* comments. This lack of effects on summarization indicated that students generally followed the explicit reviewing instructions for summarization—students were told to first summarize the main point of the paper, and on average students received one summary from each reviewer (regardless of writer ability or reviewer ability).

Praise comments were the only type of feedback that did not follow the typical pattern (overall, $M = 2.9$ comments per review; $SD = 1.2$). Not surprisingly, there was a trend that low-ability reviewers praised more than high-ability reviewers, especially on the high-ability papers. More things may impress the low-ability students because these students have more difficulty with writing.

Most importantly, significant differences were found among the *criticism* comments (see Fig. 2). On average there were three pieces of criticism per review ($SD = 1.3$). For this type of feedback, writer ability interacted with reviewer ability, $F(1, 27) = 5.39$,

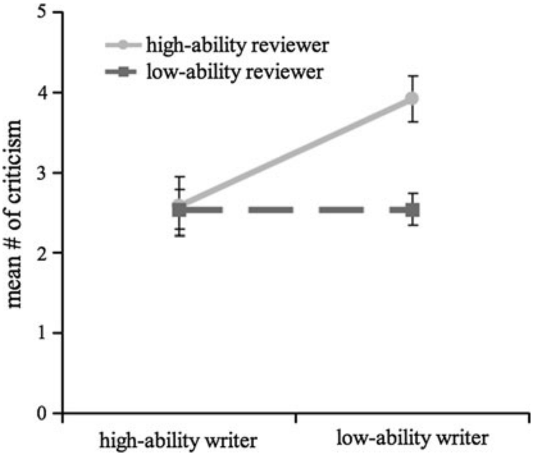


Fig. 2 Significant interactions from the writer's perspective on mean number of criticism comments

$p = .03$. The low-ability writers benefited more from the high-ability reviewers by receiving more criticism, $t(14) = 3.79$, $p < .01$; $d = 1.5$. The high-ability writers, however, received the same amount of criticism comments from high-ability and low-ability reviewers, $t(13) < 1$, $p = .93$.

Type of criticism

Since there were differences in the amount of criticism, it was important to examine further whether the increase was broadly across all criticism types or whether it was specific to particular types of criticism. For example, each criticism comment could have included *problems* and/or *solutions*. On average, two problems (i.e., *problems* or *both*; $SD = 1.1$) and two solutions (i.e., *solutions* or *both*; $SD = 1.1$) were described per review. A significant interaction was found only for feedback that described what was wrong with the paper, $F(1, 28) = 14.07$, $p < .01$ (see Fig. 3). Low-ability writers received more problems from high-ability reviewers than low-ability reviewers, $t(14) = 4.27$, $p < .01$; $d = 1.5$, but the high-ability writers received the same amount of problems from both high-ability and low-ability reviewers, $t(14) < 1$, $p = .46$.

This interaction pattern for *problems* was similar to the interaction found with *criticism* overall. While the interaction was not statistically significant for *solutions*, there was a non-significant trend that followed the same pattern. Low-ability writers received non-significantly more solutions from high-ability reviewers than low-ability reviewers, $t(14) = 1.70$, $p = .11$; $d = .7$, but the high-ability writers received the same amount of solutions from both high-ability and low-ability reviewers, $t(15) < 1$, $p = .79$.

Focus of the feedback

Another important feature of criticism is the focus of the feedback (i.e., *high prose*, *low prose*, or *substance*), where writing ability may play an even stronger role in what kinds of errors were noticed or what kinds of solutions were offered. Because similar results were found for *solutions* and *problems*, the focus of the feedback was collapsed across problem

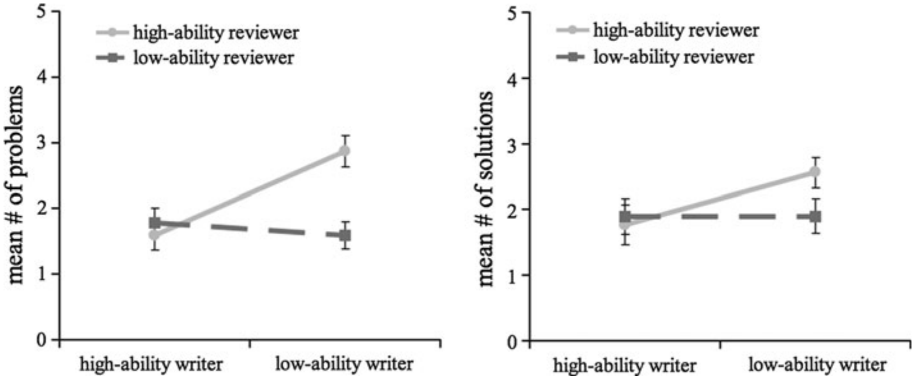


Fig. 3 Significant interactions from the writer's perspective on mean number of problems and solutions

and solution types of criticism (overall, high prose: $M = 1.4$ comments per review; $SD = 1.0$; low prose: $M = 1.1$ comments per review; $SD = 1.0$; substance: $M = 1.3$ comments per review; $SD = .9$).

Contrary to the general pattern of interaction effects, the only significant difference for feedback that focused on *high prose* was a main effect for writer ability—low-ability writers received more high prose feedback than high-ability writers, $F(1, 27) = 11.27$, $p < .01$. The main effect result was not surprising since low-ability writers are likely to make more writing errors, especially at a higher level (e.g., organization, making clear transitions, supporting arguments). More interestingly, though, the low-ability reviewers were equally likely to identify these issues in others' writing.

By contrast, significant interactions were found for both *low prose* and *substance* comments, $F(1, 28) = 5.60$, $p = .03$; $F(1, 27) = 7.07$, $p = .01$, respectively (see Fig. 4). First, high-ability writers were equally likely to receive feedback that focused on low prose issues from high-ability reviewers and low-ability reviewers, $t(13) < 1$, $p = .63$; but the low-ability writers received more feedback that focused on low prose issues from the high-ability reviewers than the low-ability reviewers, $t(13) = 2.99$, $p = .01$, $d = .9$. From the perspective that high ability writers are likely better able to detect high prose writing issues, it is perhaps surprising that high-ability writers made more low prose comments. Perhaps the high-ability reviewers were more offended by the low-prose errors that were noticed.

Turning to substance comments, the low-ability writers received more feedback that focused on substance issues from high-ability reviewers than the low-ability reviewers, $t(13) = 2.62$, $p = .02$, $d = 1.1$. The high-ability writers received the same amount of feedback that focused on substance from both high-ability reviewers and low-ability reviewers, $t(14) = 1.11$, $p = .29$. Thus, low-ability writers likely had more substance problems, but high-ability reviewers were better able to detect these problems.

Overall, only the pattern of these interactions for *substance* and *low prose* was similar to the interactions found for overall *criticism* and *problems*; that is, the overall writer ability

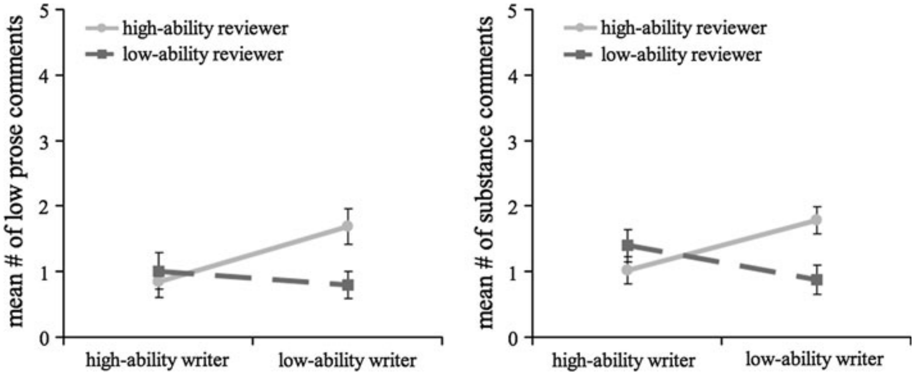


Fig. 4 Significant interactions from the writer's perspective on mean number of low prose and substance comments

by reviewer ability interaction effect appeared to be driven by differences in *low prose* and *substance* feedback.

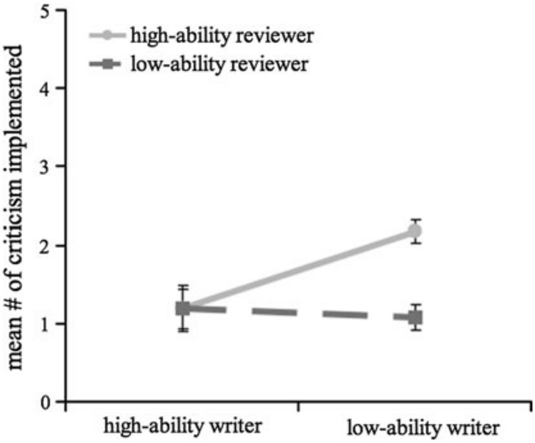
Implementation

High-ability and low-ability reviewers provided different amounts of criticism; specifically high-ability reviewers identified more low prose and substance issues in the low-ability writers' papers. As feedback does not always lead to change, one may wonder whether these differences affected the amount of revision a writer did with the feedback. On average, students implemented about half the criticisms received ($M = 1.5$ implemented revisions; $SD = 1.0$). In addition, implementation rates in the high-ability writer case provide an important indirect measure of feedback quality—if low-ability reviewers provided lower quality feedback than high-ability reviewers, the high-ability writers (who were more capable judges of suggested revisions) may be less likely to implement low-ability feedback than high-ability feedback.

Overall, writer ability interacted with reviewer ability, $F(1, 29) = 5.05$, $p = .03$ (see Fig. 5). The low-ability writers were more likely to implement the feedback from high-ability reviewers than low-ability reviewers, $t(14) = 4.47$, $p < .01$; $d = 1.7$. High-ability writers implemented the same amount of feedback from low-ability reviewers and high-ability reviewers, $t(15) < 1$, $p = .98$. This result was especially important because it suggested that the quality of feedback from low-ability and high-ability reviewers was the same.

Importantly, the patterns from the focus of the feedback were mirrored in the implementation rates for *high prose*, *low prose*, and *substance* comments (overall, high prose: $M = .6$; $SD = .6$; low prose: $M = .7$; $SD = .8$; substance: $M = .8$; $SD = .7$)—that is, the amount of feedback provided closely predicted the amount of implemented changes across writer and reviewer types. The only significant effect for the amount of implemented high prose comments was the main effect of writer ability, $F(1, 29) = 5.43$, $p = .03$. This pattern matched what was found for type of comments provided—low-ability writers

Fig. 5 Significant interactions from the writer's perspective on mean number of implemented criticism comments



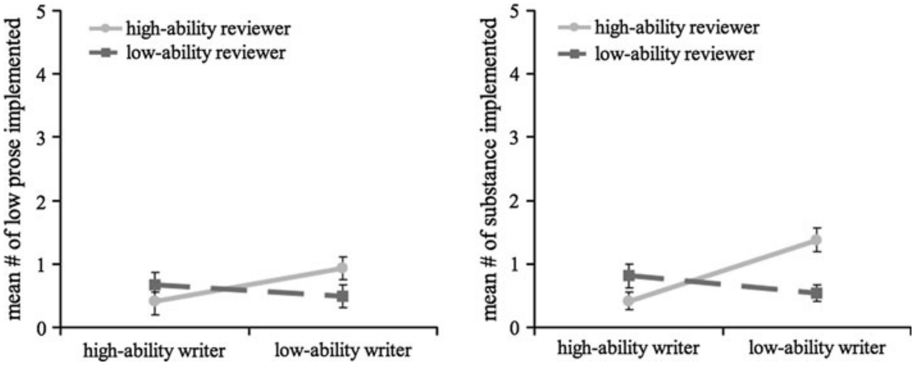


Fig. 6 Significant interactions from the writer's perspective on mean number of implemented low prose and substance comments

received more high prose comments than high-ability writers, therefore they also implemented more high prose feedback than the high-ability writers. By contrast, there were significant interactions for both low-prose implementation and substance implementation, $F(1, 28) = 4.04$, $p = .05$;, $F(1, 30) = 15.68$, $p < .01$, respectively (see Fig. 6). There were no differences in the amount of low prose and substance feedback that a high-ability writer received, so there were also no differences in the amount of implemented low prose and substance comments, $t(15) = 1.01, p = .33$; $t(15) = 1.56, p = .14$, respectively. Low-ability writers, however, received more low prose feedback from high-ability reviewers than low-ability reviewers, and the low-ability writers also implemented marginally more low prose comments from high-ability reviewers than low ability reviewers, $t(13) = 1.94$, $p = .08$, $d = .6$. The low-ability writers also received more substance feedback from high-ability reviewers than low-ability reviewers, and they implemented more substance comments from high-ability reviewers than low ability reviewers, $t(15) = 4.74$, $p < .01$, $d = 1.3$.

Again, the pattern of these interactions was similar to the ones found with the *criticism, problems*, and feedback that focused on *low prose* and *substance*. Therefore, the effect of criticism (specifically low prose and substance issues) extended to implementation overall, as well as to implementation of low prose and substance comments.

General discussion

Summary of results

Three patterns of results were found. First, high-ability reviewers provided more feedback (which was more likely to be implemented) than low-ability reviewers. Second, low-ability writers received more feedback and implemented more feedback than the high-ability writers. Most interesting were the interactions, which appeared to drive the main effects. A common pattern was found to extend into the deeper analyses of the feedback and into implementation. High-ability writers benefited equally from low-ability or high-ability reviewers, but the low-ability writers benefited more from the high-ability reviewers. The low-ability writers received more criticism in the form of problems that focused on the low-prose and substance of the paper from high-ability reviewers, which led to more feedback being implemented from the high-ability reviewers; specifically, more comments that focused on low-prose and substance were implemented. All of these benefits for low-ability writers were large effect sizes.

One pragmatic expectation in writer/reviewer ability matching situations might be a blind-leading-the-blind outcome (i.e., the low/low combination should systematically involve the lowest performance). However, across all of the interaction graphs, the low-ability writer/low-ability reviewer condition performed similarly to at least one other pairing, and some times two other pairings. Thus, the 'blind-leading-the-blind' phenomenon was not evident in this context. Instead, it is generally the case that the high-ability reviewer/low-ability writer is the best situation.

What about quality?

Some may wonder about final paper quality. In this context, students typically received comments from both high-ability and low-ability reviewers, so it was unclear whether overall quality improvements were a result from one reviewer ability or the other. In addition, analyzing the quality of writing can be complicated to measure because different audiences may value various aspects of writing more than other audiences. That is, undergraduate students and more expert writers (e.g., faculty or graduate students) may assume that different aspects of writing have a bigger impact on the quality. These differences are important when considering that the writing assignment named undergraduate students as the intended audience, and thus faculty or graduate students may not be the best judges of writing quality here.

Because the main goal of this paper was to understand how the diversity of student ability affects the process of peer-review rather than just the outcome, we used a naturalistic design (i.e., a course that utilized random assignment of reviewers for the peer-review task). As a result, it was not possible to tease apart the effects of the different ability levels on the quality of the paper. Instead, we selected aspects of the process (e.g., focus of the feedback and whether it was implemented) that are likely to affect the quality of the paper. Therefore, the papers revised by low-ability students, who received and implemented more substance feedback from the high-ability students, are likely to be of higher quality than those who received feedback from other low-ability students. In addition, the quality of the revised papers by high-ability students would not be different across review sources because the high-ability students received similar types of feedback and implemented feedback equally from high-ability and low-ability students reviewers.

Another issue of quality that needs to be investigated is the quality of the feedback provided by the students within a given type of feedback (e.g., the value of a high prose suggestion, or the accuracy of a named substance problem). This issue is particularly complicated because it is unclear who would be the most appropriate judge—that is, an expert versus a student. Ultimately, for a student to do anything with the feedback, he or she must first understand it and perceive it as helpful. A recent finding suggested that experts might overlook feedback that students would find helpful (Cho et al. 2008). Experts and peers evaluated the helpfulness of feedback that was either provided by an expert or by a peer. Peers found both types of feedback equally helpful, whereas the experts found the expert feedback to be more helpful. Because of these differences in perceived helpfulness between peers and experts, it is unclear who would be the appropriate judge of quality of feedback.

Despite these difficulties, there was some indirect evidence that the low-ability students produced comments that were of equal quality to the high-ability students—that is, high-ability students implemented the low-ability students' feedback just as often as they implemented the high-ability students' feedback. A lack of a ceiling effect in high-ability students' implementation of feedback suggests that they were able to differentiate the quality of feedback, but this differentiation was not based on the ability level of the reviewer; in general, the feedback from low-ability students was likely to be the same quality as the high-ability students.

Comparisons to prior literature

Based on prior research, it is unclear how to best assign peer assessment groups because the empirical evidence has been mixed. We considered different predictions depending upon which literature was examined. According to the cooperative learning literature, students should benefit the most from feedback provided by peers of similar ability, and this effect was likely to be moderated by the student's ability. That is, low-ability students were likely to benefit from feedback provided by high-ability peers, but high-ability students may benefit equally from both their high-ability and low-ability peers. According to the peer tutoring literature, students would likely benefit the most from feedback provided by peers of different ability. Finally, according to the writing literature, students may not benefit more from same ability or different ability peers because the two groups use very similar commenting styles. Based on the findings of the current study, the effects of ability on peer assessment seem to be most similar to the effects found in the cooperative learning literature (Lou et al. 1996)—that is, low-ability students benefited more from working with high-ability students than low-ability students, and high-ability students benefited equally from working with high-ability and low-ability students.

In addition, the features examined in the current study were chosen because they were identified as important in Nelson and Schunn (2009). They found that feedback with summary statements were likely to increase a student's understanding of the problem, and when a student understood the problem, he/she was more likely to implement the comments. In the current study, we did not find an effect of students' ability on the amount of summary comments received. Therefore, while summary comments may be important for revision tasks, they do not seem to be differentially important. Further, Nelson and Schunn found that comments with solutions were more likely to be implemented. In the current study, students' ability only marginally affected the amount of solutions received—that is, low-ability writers received non-significantly more solutions from high-ability reviewers than low-ability reviewers, but the high-ability writers received the same amount of solutions from both high-ability and low-ability reviewers. This result indicates that comments from high-ability reviewers and low-ability reviewers may be equally helpful for high-ability writers because they may receive the same amount of solutions from both sources; however, comments may be more effective for low-writers if provided by high-ability reviewers because they may include more solutions.

Caveats and future directions

First, the assignment context of this study could have influenced these results. Specifically, the focus of the feedback may have been influenced by the detailed rubric that focused on high-level writing issues. The three dimensions chosen provided scaffolding to students so that they would focus on both high prose issues (such as having clear main ideas and smooth transitions between ideas) and content issues (such as providing appropriate support and considering possible counter-arguments). Without this kind of support, novice writers tend to focus more on lower-level issues (Wallace and Hayes 1991). If a detailed rubric was not offered to the students, two possible results may occur. High-ability reviewers may focus on high-level issues while the low-ability reviewers may focus on low-level issues, or there may be no differences between high-ability and low-ability reviewers—all students may focus equally on low-level issues. Also, the assignment called for end comments (i.e., comments written outside of the paper) instead of marginalia (i.e., comments written in the margins of the paper), which would likely affect the focus of the comments. End comments are more likely to focus on high prose issues, whereas marginalia may more likely focus on low prose issues.

Another potential caveat involves range restriction. Because the students in this study attend a top-tier university, even the low-ability students are not necessarily that low—the average SAT verbal score for low-ability students was 536 (the US national SAT mean varies from year to year, but is typically between 500 and 510). Consequently, the magnitude of reviewer and writer ability interactions may be underestimated in this study; larger effects may occur in lower-tier universities where students with lower SAT scores are accepted or secondary schools where the low-ability students may include students with learning disabilities. Future research should examine a broader range of student abilities, considering both higher versus lower ability differences within a lower overall ability context (e.g., as would be the case at a less selective university) as well as high/low pairings in more diverse ability contexts.

Low-ability writers appeared to benefit the most from high-ability reviewers. However, we were not able to determine how the differences we found affected the final quality of the paper or whether students learned more from the high-ability reviewers (i.e., would the low-ability students do better on a second paper after revising a paper based on high-ability reviewers' feedback?). To properly answer these questions, future research should involve randomly assigning high-ability and low-ability students several reviewers of the same ability (i.e., all high-ability reviewers or all low-ability reviewers.

In the current study, writer ability was measured using an indirect measure (i.e., self-reported SAT verbal). For research purposes, future studies should involve using more direct measures of writer ability. For pragmatic purposes, these indirect measures have great value (i.e., easy to collect at the beginning of the semester). Therefore, future research should validate which indirect measures are most appropriate.

Many questions still remain regarding the quality of the feedback provided by peers. Are all students capable of constructing useful feedback? How closely to students and instructors agree on what constitutes high quality feedback? Do students focus on the important features like summarization, localization, and solutions (Nelson and Schunn 2009) when rating the quality of feedback?

Finally, the participants in the current study not only received feedback, but also provided feedback. Recent research has demonstrated that there are significant benefits to providing feedback (Cho and Cho 2011). While students' revisions in the current study might be affected by providing comments to their peers, the focus of the analyses were on what kinds of comments were received and whether students' implemented these comments. Because the issue of providing feedback is very important, future research should further examine whether these benefits depend on the writer's ability level and/or the reviewer's ability level.

Implications for classroom practice

Low-ability writers appeared to benefit the most from high-ability reviewers. The low-ability writers received more feedback that identified problems with the substance of the paper, and they implemented more feedback from the high-ability reviewers. High-ability writers appeared to benefit equally from both the high-ability and the low-ability reviewers. Therefore, students appeared to benefit the most from peer assessment when they were assigned to groups with peers of dissimilar ability. In addition, self-reported SAT verbal scores appeared to be reasonably easy to obtain surrogates for writing ability that instructors could use to create these groups, although researchers may wish to use more direct writing ability measures.

Appendix

See Tables 2 and 3.

Table 2 Coding scheme

	Definition	Example
Type of feedback (kappa = .97)		
Summary	Summarization of the main points of the paper	The main argument [sic.] is that stereotypes and emotions are (somehow) related
Praise	Complimentary comment or identifying a positive feature in the paper	I like your use of quotes throughout the paper
Criticism	A comment in which a problem and/or a solution was described	Also, some points were not necessary, such as your math example
Type of criticism (kappa = .86)		
Problem	Something that needs to be fixed is explicitly identified	The author also did not consider the possibility of a counter-argument
Solution	A possible improvement is explicitly offered	Also, I would cite your sources a little more thoroughly
Both	The reviewer is provides both a problem and solution	The arguments however were not exactly clear. Try to make them clearer
Focus of the feedback (problem: kappa = .70; solution: kappa = .74)		
High prose	A comment focusing on a reviewing dimension (i.e., flow, logic, insight)	You did not use any citations to back up your claims
Low prose	A comment focusing on low-level writing issues (e.g., grammar, word choice)	The use of "I" throughout the paper is distracting
Substance	A comment focusing on issues that only someone with content knowledge could fix	Para 1 pg 2 What does "accidental pieces of intentional behavior mean?"

Table 3 Descriptive & inferential statistics

	Reviewer ability	Writer ability						ANOVA			*t* test	
		High			Low			Writer ability	Reviewer ability	Interaction	For high-ability writers	For low-ability writers
		N	*M*	*SD*	*N*	*M*	*SD*					
Amount of feedback												
# of segments	High	14	6.5	1.6	14	7.3	1.4	$p = .31$	$p = .60$	$p = .28$	$p = .73, d = -.2$	$p = .19, d = .6$
	Low	14	6.8	1.4	14	6.6	1.1					
Type of feedback												
Summary	High	14	.8	.4	14	.7	.4	$p = .98$	$p = .99$	$p = .52$	$p = .70, d = .2$	$p = .58, d = -.2$
	Low	14	.7	.5	14	.8	.5					
Praise	High	13	2.6	1.1	16	2.6	1.3	$p = .38$	$p = .12$	$p = .48$	$p = .15, d = -.7$	$p = .52, d = -.2$
	Low	13	3.4	1.1	16	2.9	1.1					
Criticism	High	14	2.6	1.3	15	3.9	1.1	$p = .02$	$p = .02$	$p = .03$	$p = .93, d = .0$	$p < .01, d = 1.5$
	Low	14	2.5	.9	15	2.5	.7					
Type of criticism												
Problems + both	High	15	1.6	.9	15	2.9	.9	$p = .04$	$p = .01$	$p = .001$	$p = .46, d = -.2$	$p < .01, d = 1.5$
	Low	15	1.8	.9	15	1.6	.8					
Solutions + both	High	16	1.8	1.2	15	2.6	.9	$p = .10$	$p = .39$	$p = .21$	$p = .79, d = -.1$	$p = .11, d = .7$
	Low	16	1.9	1.1	15	1.9	1.0					
Focus of criticism												
High prose	High	15	1.1	.9	14	2.0	.8	$p = .002$	$p = .51$	$p = .67$	$p = .86, d = .1$	$p = .50, d = .3$
	Low	15	1.0	.9	14	1.7	1.1					
Low prose	High	14	.8	.8	16	1.7	1.1	$p = .25$	$p = .12$	$p = .03$	$p = .63, d = -.2$	$p = .01, d = .9$
	Low	14	1.0	1.0	16	.8	.8					
Substance	High	15	1.0	.8	14	1.8	.8	$p = .55$	$p = .27$	$p = .01$	$p = .29, d = -.4$	$p = .02, d = 1.1$
	Low	15	1.4	.9	14	.9	.8					

Table 3 continued

	Reviewer ability	Writer ability						ANOVA			*t* test	
		High			Low			Writer ability	Reviewer ability	Interaction	For high-ability writers	For low-ability writers
		N	*M*	*SD*	*N*	*M*	*SD*					
Implementation												
Total implemented	High	16	1.2	1.2	15	2.2	.6	$p = .06$	$p = .03$	$p = .03$	$p = .98$, $d = .0$	$p < .01$, $d = 1.7$
	Low	16	1.2	1.0	15	1.1	.7					
High prose	High	15	.4	.6	16	.9	.7	$p = .03$	$p = .47$	$p = .59$	$p = .89$, $d = .0$	$p = .37$, $d = .3$
	Low	15	.4	.5	16	.7	.6					
Low prose	High	16	.4	.6	14	.9	.7	$p = .36$	$p = .61$	$p = .05$	$p = .33$, $d = -.4$	$p = .08$, $d = .6$
	Low	16	.7	.8	14	.5	.7					
Substance	High	16	.4	.6	16	1.4	.7	$p = .06$	$p = .18$	$p = .00$	$p = .14$, $d = -.6$	$p < .01$, $d = 1.3$
	Low	16	.8	.8	16	.5	.5					

References

Arum, R., & Roksa, J. (2011). *Academically adrift: Limited learning on college campuses*. Chicago: University of Chicago Press.

Bok, D. (2006). *Our underachieving colleges: A candid look at how much students learn and why they should be learning more*. Princeton, NJ: Princeton University Press.

Cho, Y., & Cho, K. (2011). Peer reviewers learn from giving comments. *Instructional Science, 39*(5), 629–643.

Cho, K., Chung, T. R., King, W. R., & Schunn, C. D. (2008). Peer-based computer-supported knowledge refinement: An empirical investigation. *Communications of the ACM, 51*(3), 83–88.

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.

Cobb, K. L. (1999). Interactive videodisc instruction with undergraduate nursing students using cooperative learning strategies. *Computers in Nursing, 17*(2), 89–96.

Cohen, P. A., Kulik, J. A., & Kulik, C.-L. C. (1982). Educational outcomes of tutoring: A meta-analysis of findings. *American Educational Research Journal, 19*(2), 237–248.

Day, E. A., Arthur, W., Bell, S. T., Edwards, B. D., Bennett, W., Mendoza, J. L., et al. (2005). Ability-based pairing strategies in the team-based training of a complex skill: Does the intelligence of your training partner matter? *Intelligence, 33*(1), 39–65.

De Lisi, R., & Golbeck, S. L. (1999). Implications of Piagetian theory for peer learning. In A. M. O'Donnell & A. King (Eds.), *Cognitive perspectives on peer learning* (pp. 3–37). Mahwah, NJ: Lawrence Erlbaum Associates.

Goethals, G. R. (2001). *Peer effects, gender, and intellectual performance among students at a highly selective college: A social comparison of abilities analysis*. Discussion Paper. Document Number.

Goethals, G. R. (2002). *Social comparison and peer effects at an elite college*. Discussion Paper, Document Number.

Gouli, E., Gogoulou, A., & Grigoriadou, M. (2008). Supporting self-, peer-, and collaborative-assessment in e-learning: The case of PEer and Collaborative ASSessment Environment (PECASSE). *Journal of Interactive Learning Research, 19*(4), 615–647.

Jehn, K. A., Northcraft, G. B., & Neale, M. A. (1999). Why differences make a difference: A field study of diversity, conflict, and performance in workgroups. *Administrative Science Quarterly, 44*(4), 741–763.

Lin, S. S., Liu, E. Z., & Yuan, S. M. (2001). Web-based peer assessment: Feedback for students with various thinking-styles. *Journal of Computer Assisted Learning, 17*(4), 420–432.

Lou, Y., Abrami, P. C., Spence, J. C., Poulsen, C., Chambers, B., & d'Apollonia, S. (1996). Within-class grouping: A meta-analysis. *Review of Educational Research, 66*(4), 423–458.

Miller, G., & Polito, T. (1999). The effect of cooperative learning team compositions on selected learner outcomes. *Journal of Agricultural Education, 40*(1), 66–73.

Moshman, M., & Geil, D. (1998). Collaborative reasoning: Evidence for collective rationality. *Thinking & Reasoning, 4*(3), 231–248.

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.

Patchan, M. M., Charney, D., & Schunn, C. D. (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), 124–152.

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.

Topping, K. J. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research, 68*(3), 249–276.

Topping, K. J. (2005). Trends in peer learning. *Educational Psychology, 25*(6), 631–645.

Topping, K. J., & Ehly, S. (1998). *Peer-assisted learning*. Mahwah, NJ: Lawrence Erlbaum Associates.

Tutty, J. I., & Klein, J. D. (2008). Computer-mediated instruction: A comparison of online and face-to-face collaboration. *Educational Technology Research and Development, 56*(2), 101–124.

Vygotsky, L. S. (1978). *Mind in society*. Cambridge, MA: Harvard University Press.

Wallace, D. L., & Hayes, J. R. (1991). Redefining revision for freshmen. *Research in the Teaching of English, 25*(1), 54–66.

Watson, S. B., & Marshall, J. E. (1995). Effects of cooperative incentives and heterogeneous arrangement on achievement and interaction of cooperative learning groups in a college life-science course. *Journal of Research in Science Teaching, 32*(3), 291–299.

Webb, N. M., Troper, J. D., & Fall, R. (1995). Constructive activity and learning in collaborative small groups. *Journal of Educational Psychology, 87*(3), 406–423.

Wooley, R. S., Was, C., Schunn, C. D., & Dalton, D. (2008). *The effects of feedback elaboration on the giver of feedback*. Paper presented at the Cognitive Science