# Understanding the Effects of Receiving Peer Feedback for Text Revision: Relations between Author and Reviewer Ability

Melissa M. Patchan and Christian D. Schunn

University of Pittsburgh, PA | USA

Abstract: Peer assessment is a technique with many possible benefits for instruction across the curriculum. However, the value obtained from receiving peer feedback may critically depend upon the relative abilities of the author and the reviewer. We develop a new model of such relative ability effects on peer assessment based on the well-supported Flower and Hayes model of revision processes. To test this model across the stages of peer assessment from initial text quality, reviewing content, revision amount, and revision quality, 189 undergraduate students in a large, introductory course context were randomly assigned to consistently receive feedback from higher-ability or lower-ability peers. Overall, there were few main effects of author ability or reviewer ability. Instead, as predicted, there were many interactions between the two factors, suggesting the new model is useful for understanding ability factors in peer assessment. Often lower-ability writers benefitted more from receiving feedback from lower-ability reviewers, while higher-ability writers benefitted equally from receiving feedback from lower-ability and higher-ability reviewers. This result leads to the practical recommendation of grouping students by ability during peer assessment, contrary to student beliefs that only feedback from high ability peers is worthwhile.

Keywords: peer assessment; writing ability; reviewing ability; peer feedback; revisions

journal of WRITING RESEARCH

Patchan, M. M., & Schunn, C. D. (2016). Understanding the effects of receiving peer feedback for text revision: Relations between author and reviewer ability. *Journal of Writing Research, 8*(2), 227-265. doi: 10.17239/jowr-2016.08.02.03

Contact: Melissa M. Patchan, University of Pittsburgh, 4033 Forbes Towers, Pittsburgh, PA 15260 | USA - mpatchan@pitt.edu.

Many students graduate high schools in the United States without achieving a proficient level of writing skills (National Center for Education Statistics, 2012). Not surprisingly, college students' writing ability often does not meet educators' expectations. This problem is further compounded by the numerous large, content classes (i.e., more than 75 students), which make grading and providing adequate feedback on writing assignments very difficult for instructors (Arum & Roksa, 2010; Bok, 2006). In short, instructors rarely assign any substantive writing assignments or do not include multi-draft writing assignments, in which students read and act upon received feedback, because instructors do not have the time to provide feedback. To address these problems, instructors utilize peer assessment as a means to incorporate writing assignments and individualized timely feedback in large classes.

Peer assessment (often also called peer review) is the quantitative evaluation and qualitative feedback 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. Researchers have advocated peer assessment for more than four decades (Bruffee, 1980; Elbow, 1973; Moffett, 1968). In small composition classes, students will often exchange documents and have pair or group-based face-to-face conversations about their documents with a primary emphasis on qualitative feedback (Chang, 2012; DiPardo & Freedman, 1988). In larger classes or when writing occurs outside of composition classes, peer assessment takes a form that is more like professional journal reviewing practices: quantitative and qualitative assessment, asynchronous, and written-based comments. These two kinds of assessment are quite different in character, especially in terms of important pragmatics such as the possibility of anonymity and the amount of interaction and follow-up. Because peer assessment in the written, asynchronous form is growing but still far from universally implemented, we focus on that form of peer assessment.

Research on such peer assessments has found that, in general, peers are capable of providing valid ratings (for a review, see Falchikov & Goldfinch, 2000; see also Bouzidi & Jaillet, 2009; Cho, Schunn, & Wilson, 2006; Panadero, Romero, & Strijbos, 2013; Topping & Fisher, 2003; Tseng & Tsai, 2007), and their feedback is usually just as effective as an instructor's feedback in helping students improve their drafts (for a review, see Topping, 2005; see also Gielen, Tops, Dochy, Onghena, & Smeets, 2010) and sometimes more effective (Cho & MacArthur, 2011; Cho & Schunn, 2007; Hartberg, Gunersel, Simspon, & Balester, 2008). Further, participating in peer assessment improves writing ability (Topping, 1998).

The overall benefits of peer assessment likely come from several sources. First, peers may be better able to articulate feedback in terms that are understandable (Cho & Schunn, 2007; Patchan, Schunn, & Clark, 2011). Second, multi-peer assessment can provide more total feedback than an over-taxed instructor (Cho, Schunn, & Charney, 2006; Patchan, Charney, & Schunn, 2009; Patchan, Schunn, & Clark, 2011), more persuasive feedback when multiple reviewers note the same problems, and feedback representing more diverse audience perspectives (DiPardo & Freedman, 1988;

Mumford, 1983). Third, the act of providing feedback to others is itself a learning opportunity (Lundstrom & Baker, 2009; Topping, Dehkinet, Blanch, Corcelles, & Duran, 2013), including learning from seeing models of effective and ineffective writing (Charney & Carlson, 1995) as well as practicing revision strategies (Patchan & Schunn, 2015). In other words, students can be conceived as practicing revision through the act of proposing revisions to their peers.

Despite this evidence that peer assessment can benefit students, some instructors are reluctant to use peer assessment, and students are often skeptical of the usefulness of receiving peer feedback because they are concerned that not all peers are capable of helping them (Kaufman & Schunn, 2011). This skepticism was clearly articulated in several student interview responses from a previous pilot study. For example, when asked whether they felt that receiving peer feedback was useful, an exemplar student response was, "I think it can be but it depends how knowledgeable the peers are." Another student asked the same question elaborated on this point by saying, "I think it is if they…if their skill level is more advanced than mine, or at least very similar. […] I don't think it would be as much if the person…if their writing wasn't at the level that mine was, then no I don't think that they can really provide feedback that's going to advance."

As these examples demonstrate, students sometimes have negative beliefs about what happens in peer assessment, which could impact whether instructors feel comfortable utilizing it in their classroom. Researchers have wondered whether some kind of matching by ability is needed (Topping, 1998), and some researchers have built tools that automatically match students as reviewers based on ability (Crespo Garcia & Pardo, 2010; Crespo, Pardo, & Delgado Kloos, 2004; Gehringer, 2001; Giannoukos et al., 2010; Hsiao & Brusilovsky, 2008). However, before more such interventions are developed, the question remains whether student, instructor, and system designer suppositions are accurate. More information about the kinds of feedback received by students in peer assessment and how this feedback varies by ability is needed. Therefore, the goal of the current study was to document the relationship between author ability and reviewer ability and its effect on peer assessments received by students. To understand how peer feedback might vary in important ways, we review accounts of how writing ability varies and then propose how that writing-ability variation should relate to peer assessment variation.

### 1.1 A Theoretical Model of Ability Differences in Revision Processes

Flower and Hayes developed a well-cited Cognitive Process Theory of writing that involved three cognitive processes: planning, translating, and reviewing, which were controlled by a monitor (Flower & Hayes, 1981). The original theory and its more recent updates continue to be highly influential in understanding ongoing issues within writing research, including the effect of planning and environmental factors on metacognitive processes of second-language writers (Ong, 2014), how to support planning with the use of outlines (de Smet, Brand-Gruwel, Broekkamp, & Kirschner,

2012), and how to support text production and revision in novice writers (Pifarré & Fisher, 2011). Most relevant to peer assessment is the reviewing process, which was further elaborated as comprising several cognitive processes, including problem detection, problem diagnosis, and strategy selection (Hayes, Flower, Schriver, Stratman, & Carey, 1987). This original model has been updated to better understand cognition and affect in writing (Hayes, 1996; Hayes, 2012). One surprising change was the removal of the reviewing process, for which Hayes explained that revision was better thought of as a special application of the updated model rather than its own process (Hayes, 2012). Despite the lack of a reviewing process in the updated model, problem detection, problem diagnosis, and strategy selection are still relevant activities in revision and peer assessment.

Reviewing begins with text reading, and students of lower writing ability may have weaker reading skills, which limit their ability to detect problems in others' writing (Hayes, 1996). Problem detection is the reviewing process in which a writer perceives differences between the text produced so far and the intended text (Hayes et al., 1987; Scardamalia, Bereiter, & Steinbach, 1984). Problem detection is an essential process because it must occur before any revisions can be made. In studies of differences of expert versus novice writers, higher-ability writers were found to detect more problems than lower-ability writers. Furthermore, they are much more likely to detect global problems (Fitzgerald, 1987; Hayes et al., 1987; Scardamalia & Bereiter, 1983; Sommers, 1980). There are two possible explanations for this difference. First, higher-ability writers have more knowledge about problems that occur in writing, and they have richer knowledge for each type of problem. This extensive knowledge can help higher-ability writers not only detect more problems but also do so more easily. In addition to knowing more about writing problems, higher-ability writers also approach revision with a different, more appropriate, task definition—that is, their primary goal is to rethink, which is more likely to result in improvements to the overall text (Scardamalia & Bereiter, 1987; Scardamalia & Bereiter, 1983). Lower-ability writers, on the other hand, tend to fixate on repairing errors, which limits their revisions to very localized changes that do very little to improve the overall text.

Problem diagnosis involves creating a representation of the problems detected (Scardamalia & Bereiter, 1984; Hayes et al., 1987). To be useful, this representation must provide enough information for the writer to determine an appropriate reaction. A diagnosis can vary in the amount and level of explicitness, with well-defined representations (i.e., knowledge about the problem including the location and the cause, which often leads to a specific solution to the problem) at the high end of the continuum and ill-defined representations (i.e., only knowing that something does not sound right) at the low end of the spectrum. Whether a writer diagnoses a problem is dependent on what problems are detected. Stronger writers have more existing schemas for particular kinds of writing issues (Hayes, 1996). Therefore, it is not surprising that higher-ability writers diagnose more problems overall than lower-ability writers, and in particular, their diagnoses tend to be more global. In addition, because higher-ability

writers have more information about problems based on their more elaborate writing schemas, they tend to provide more elaborate diagnoses.

Strategy selection involves reacting to a detected problem (Hayes et al., 1987). This process involves both decision-making skills as well as problem solving skills. First, writers must decide on which problems to solve and which strategy would be most effective. When problems are ill-defined or the most appropriate strategy is not obvious, the writer must utilize a search strategy to discover new solutions to the problem at hand. As a result, the quality of the solutions applied varies by writing ability. Similar to problem diagnosis, strategy selection is also dependent on problem detection. Higher-ability writers make more revisions than lower-ability writers, and specifically, they make significantly more global revisions. Higher-ability writers not only have a richer knowledge of more problems, they also have more solutions matched to those problems. This more sophisticated means-ends repertory helps higher-ability writers choose more effective solutions to the writing problems they detect.

### 1.2 Predicting Ability Effects in Peer Assessment of Writing

Beginning writers' revision processes tend to be weak, and they benefit from scaffolding (Scardamalia & Bereiter, 1983). Peer assessment can provide such scaffolding (Topping, 1998), and it is important to understand how the benefits of peer assessment may vary by student ability. Using the Hayes and Flower model and the broadly observed writer ability effects on different revision practices, we can build a theoretical model of ability effects on peer assessment. The impact of these ability differences seen in revision will likely be observed at various points in peer assessment (see Figure 1). We focus on the ability of both authors and reviewers as writers; as a short hand, we refer to *author ability* as the author's abilities in writing, and *reviewer ability* as the reviewer's ability in writing.

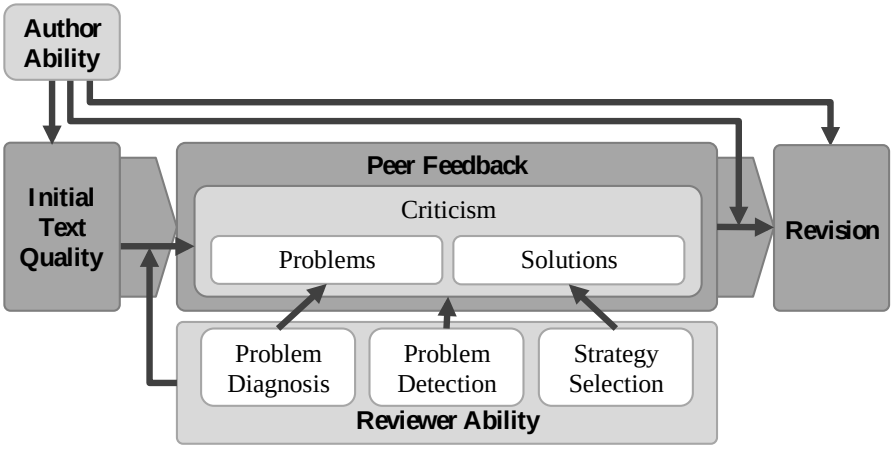



*Figure 1.* Theoretical model of ability effects on peer assessment.

In the first step of the peer assessment process, students write their initial draft. The author's ability will likely directly affect the quality of this draft; higher-ability authors can detect, diagnose, and solve more problems in their own writing—especially global problems—than lower-ability authors (Hayes et al., 1987), which will result in a higher quality draft.

For the second step, students review one or more peers' papers. This review typically includes rating the quality of the text and more importantly, providing comments for the authors to use to help improve their paper. Peer feedback is expected to be affected indirectly by the author's ability via the initial text quality (i.e., via the relative frequency of problems there are to detect and repair). Lower-quality texts would likely receive a greater number of criticism comments, as indicated in Figure 1.

Such effects have been observed in two prior studies: 1) an analysis of over 1400 comments for 24 essays written by undergraduate students enrolled in a large history course (Patchan, Charney, & Schunn, 2009) and 2) an analysis of over 1100 comments for 38 texts written by undergraduate students enrolled in a large cognitive science course (Patchan, Hawk, Stevens, & Schunn, 2013). Across both studies, lower-quality texts received more criticism comments than higher-quality texts ($d$ = 0.65 - 0.77).

Additionally, peer feedback will also likely be affected by the reviewer's ability. Students receiving feedback from higher-ability reviewers would likely receive a greater number of criticism comments, specifically comments that describe problems and offer solutions, whereas students receiving feedback from lower-ability reviewers would likely receive fewer criticism comments, and of these comments, many will not adequately describe problems or offer solutions. Furthermore, reviewer ability is expected to moderate the effect of initial text quality on peer feedback. Since higher-ability reviewers have more refined reviewing skills, they will be able to better detect the differences in text quality (i.e., more criticism for low quality papers and more praise for high quality papers). Lower-ability reviewers, whose reviewing skills are less developed, will be less able to make these distinctions.

The results from the Patchan et al. (2009; 2013) studies generally supported of these predicted main effects and interactions. Across both studies, higher-ability reviewers provided more criticism comments than lower-ability reviewers ($d$ = 0.63 - 0.75). Moreover, a significant text quality effect was found for the higher-ability reviewers—that is, higher-ability reviewers provided more criticism comments for lower-quality texts than higher-quality texts ($d$ = 0.93 - 1.09). However, a similar effect was not found for lower-ability reviewers; in general, lower-ability reviewers provided a similar amount of criticism comments to both lower-quality and higher-quality texts ($d$ = 0.00 - 0.71).

As the last step, students revise their papers based on the peer feedback they received. Similar to the initial draft, the author's ability will directly affect what revisions are made, the quality of the revisions, and ultimately the quality of the final draft. Higher-ability authors are expected to make more revisions, and they are also expected to make higher quality revisions because they are better able to choose more

effective solutions. By contrast, lower-ability writers are expected to make fewer revisions, and these revisions are expected to be of lower quality. The reviewer's ability could also indirectly affect the revisions. All authors would likely make more revisions when receiving feedback from higher-ability reviewers than lower-ability reviewers because these peers are likely to provide more criticism comments, specifically ones that offer solutions. Because higher-ability reviewers are better able to suggest effective solutions, the revisions made when implementing higher-ability reviewers' comments are expected to be of higher quality. Similar to peer feedback, there may be an interaction between author ability and reviewer ability. Because higher-ability authors have a more sophisticated means-ends repertoire, they are expected to equally address issues detected by both higher-ability reviewers and lower-ability reviewers—that is, higher-ability authors will be able to figure out how to solve a problem even if lower-ability reviewers do not provide as much information as the higher-ability reviewers. The findings from the Patchan et al. (2013) study support these predictions. There was only a marginal writer ability effect and a significant reviewer ability effect, but these effects were driven by a significant interaction—that is, lower-ability writers implemented the most feedback when receiving feedback from higher-ability reviewers, and all other writers implemented a similar number of comments. Few studies have examined the differential impact of peer feedback on the quality of the revisions (Allal, Chanquoy, & Largy, 2004).

There are a few aspects with no predictions from the model that the current study will also investigate. Although lower-ability authors have fewer solutions available to them and are less able to choose effective solutions, it is unclear whether they would be able to distinguish between more or less effective solutions when provided with both options or come up with solutions when a problem is described without one. Furthermore, it is unclear whether they will be able to successfully implement all effective solutions if unfamiliar strategies are included (i.e., they may receive suggests that are beyond their abilities to implement successfully). This gap is a critical one in understanding the effects of peer assessment on student performance and learning.

Finally, given that lower-ability writers are less likely to address global issues (Fitzgerald, 1987; Hayes et al., 1987; Sommers, 1980), these effects are expected to be especially true for comments focusing on high-level issues. Therefore, in the current study, we will separately look at comments that address high-level issues (e.g., clarity of main idea, transitions) and low-level issues (e.g., grammar).

### 1.3 Research Questions and Predictions

The goal of the current study was to document the relationship between author ability and reviewing ability and its effect on the impact of receiving peer assessments. We test a new peer assessment model that we have derived from the Flower and Hayes writing model and that already has received partial support (Allal, Chanquoy, & Largy, 2004; Patchan et al., 2009; 2013). By considering how the variations in author ability could relate to the three main steps/products of peer assessment (i.e., initial text, peer

feedback, and revisions), several research questions and predictions have emerged. For the peer feedback, 1) does the initial text quality (i.e., author ability) affect the amount and types of peer feedback received, 2) does reviewer ability affect the amount and types of peer feedback received, and 3) do author ability and reviewer ability interact? For revisions, 1) does author ability affect the amount and quality of revisions made, 2) does reviewer ability affect the amount and quality of revisions made, and 3) do author ability and reviewer ability interact? Predictions for each of these questions are organized in Table 1. As there are competing factors (e.g., how many comments received vs. how many comments were understood) and potentially complex interactions, these predictions, which were formulated by taking into account the processes in the Flower and Hayes model, need to be examined empirically.

## 2. Method

### 2.1 Overview

The current study was part of a larger study that examined multiple aspects of why students learn from peer assessment, including the relative effectiveness of different forms of peer feedback (Patchan, Schunn, & Correnti, in press) and the benefits for the reviewer of *providing* feedback (Patchan & Schunn, 2015), in contrast to the current focus on the relationship between author ability and reviewing ability on *receiving* feedback. In order to describe the extent to which author ability and reviewer ability affect peer assessment, we determined the writing ability of each participant and then manipulated from whom the participants received comments. In other words, in a 2 x 2 between-subjects design, groups of participants of higher-writing ability (i.e., high authors) or of lower-writing ability (i.e., low authors) received feedback from either only groups of peers with higher-writing ability (i.e., high reviewers) or only groups of peers with lower-writing ability (i.e., low reviewers). We examine the effects of author ability and reviewing ability on the amount, features, and focus of comments received, the implementation rate, and the revision quality were compared across the conditions.

### 2.2 Course Context

This study was conducted in an Introduction to Psychological Science course at a large, public research university in the southeast United States. The specific class and assignment context was selected to represent an authentic writing assignment that occurred in a large, content course as part of the Writing in the Discipline (WID) program. This course was a popular general education course that students commonly took to meet one of their social science requirements. In addition, it was compulsory for not only all psychology majors, but also for a number of other majors as well, including education and nursing. Because this course was very large (i.e., 838 students), three sections were offered, each taught by a different lecturer. Students were

**Table 1.** Predictions for Main Effects and Interactions of Initial Text, Peer Feedback, and Revisions

| | Initial Text | Peer Feedback | Revisions |
|---|---|---|---|
| Author Ability Effect | Higher-ability authors were expected to produce higher quality first drafts than lower-ability authors. | Lower-ability authors were expected to receive more high-level criticism than higher-ability authors. | Implementation: Higher-ability authors were expected to implement more high-level comments* than lower-ability authors.<br><br>Revision Quality: Higher-ability authors were expected to make higher quality revisions when implementing high-level comments* than lower-ability authors. |
| Reviewer Ability Effect | No reviewer effect should be present. | A greater number high-level comments* were expected from higher-ability reviewers than lower-ability reviewers. | Implementation: Authors were expected to implement more high-level comments* from higher-ability reviewers than from lower-ability reviewers.<br><br>Revision Quality: Authors were expected to make higher quality revisions when implementing high-level comments* from higher-ability reviewers than from lower-ability reviewers. |
| Interaction | No interaction should be present. | The difference in the number of high-level comments* between higher-ability authors and lower-ability authors was expected to be greater for higher-ability reviewers than lower-ability reviewers. | Implementation: Although higher-ability authors were expected to implement high-level comments* from higher-ability reviewers and lower-ability reviewers equally, lower-ability authors may implement more high-level comments* from higher-ability reviewers or lower-ability reviewers.<br><br>Revision Quality: Higher-ability authors were expected to make similar quality revisions when implementing high-level comments* from higher-ability reviewers and lower-ability reviewers equally. No prediction is made for lower-ability authors, who may either implement more high-level comments* from higher-ability reviewers or from lower-ability reviewers. |

* The term 'comments' here refers only to criticism comments: ones that include problems and/or solutions.

also required to attend one of 24 different lab sections taught by 12 graduate student teaching assistants (TAs).

## 2.3 Participants

Multiple research studies were implemented in this class. The current study used data from a random selection of 189 participants, representing the full range of writer ability, who were given documents to review from students across the ability spectrum. This random sample reflects that range and distribution of students in this class: 77% female; with students at all levels, with a predominance of less advanced students (i.e., 58% freshmen, 27% sophomores, 10% juniors, and 5% seniors); and a great variety of majors (i.e., of the declared majors: 30% social sciences, 30% natural sciences, 14% engineering, 13% education, 6% computer science, and 6% business).

Author ability and reviewer ability were then defined based on an estimate of the participants' writing ability. First, the participants' writing ability was determined by a composite of four self-reported ability measures—that is, the average *z*-scores (i.e., student's score minus group mean divided by group standard deviation) of SAT verbal[1], SAT writing, the final grades in the first and second semester composition courses[2]. This combination of measures provided a more generalizable ability measure that one can also obtain easily for future research or practical applications.

Next, a median split was used to determine which students had higher writing ability and which students had lower writing ability. Indeed, relative to the U.S. ability standards, the two groups were above and below median performance levels (The College Board, 2012). Further, given the median split approach to defining the groups and the broad ability range found in this class, there were grouping differences of 2.8 standard deviations (i.e., a very large effect size) on the composite measure, and there were also large group differences on each of the components of this composite measure (see Table 2). To further validate the composite measure, two writing experts (i.e., rhetoric graduate students with extensive writing teaching experience) rated the quality of the students' first drafts using a 5-point scale on eight dimensions focused on the flow, argument logic, and insight of the papers (see *Coding Process—Quality of Writing* for more details; rubric details are in Appendix A). The sum of all eight dimensions (i.e., out of 40 possible points) was compared between the high authors and low authors. An independent *t*-test on the sample of 189 participants revealed a significant difference in author ability: the high authors ($M$ = 17.0, $SD$ = 3.8) produced higher quality first drafts than the low authors ($M$ = 14.8, $SD$ = 3.2), $t(187) = 4.33$, $p < .001$, $d = 0.63$. Finally, students with higher writing ability were defined to be high authors and high reviewers, and students with lower writing ability were defined to be low authors and low reviewers. Although this method was not the most precise way to define author ability and reviewer ability, it was pragmatically required for creating the groups for assignment to reviewing groups for this study and in future instructional applications.
This decision decreases the power of this study, which could result in missing some relevant data patterns. However, there is little chance of making false claims, and the

overall large number of participants means that the instructionally important patterns will generally be detectable. We believe that a lower powered study was a reasonable tradeoff for higher external validity (i.e., how reviewer ability would typically be detected).

**Table 2.** Summary of Demographic and Ability Data by Writer Ability

| | High-Ability Writer | | | Low-Ability Writer | | | t-test | |
|---|---|---|---|---|---|---|---|---|
| | n | *M* | *SD* | n | *M* | *SD* | *p* | *d* |
| *Demographics* | | | | | | | | |
| gender[a] | 93 | 75% | | 95 | 78% | | .67 | |
| year in school[b] | 93 | 84% | | 96 | 82% | | .77 | |
| age | 93 | 18.8 | 1.5 | 96 | 19.0 | 1.8 | .45 | |
| *Ability Measures* | | | | | | | | |
| writing ability z-score | 93 | 0.64 | 0.49 | 96 | -0.70 | 0.47 | < .0001 | 2.8 |
| SAT verbal | 79 | 599 | 55 | 70 | 489 | 55 | < .0001 | 2.0 |
| SAT writing | 76 | 592 | 63 | 69 | 494 | 58 | < .0001 | 1.6 |
| 1st semester grade[c] | 70 | 4.4 | 0.7 | 65 | 3.3 | 0.6 | < .0001 | 1.7 |
| 2nd semester grade[c] | 49 | 4.2 | 0.7 | 51 | 3.1 | 0.6 | < .0001 | 1.7 |

[a] % female

[b] % freshman + sophomore

[c] Composition grades were coded on a 5-point scale: 5 – placed out; 4 – A, 3 – B, 2 – C, 1 – D or below. Missing data points included participants who did not take the composition course because it was not a required course ($n_{2nd\ semester} = 1$) and participants who were currently taking the course or will take it in the future ($n_{1st\ semester} = 54$; $n_{2nd\ semester} = 88$).

## 2.4 Procedure

Participants completed three main tasks: 1) wrote a first draft, 2) reviewed peers' texts, and 3) revised own text based on peer feedback. At the end of the first month of the semester, participants had one week to write their first draft and submit it online using the web-based peer assessment functions of turnitin.com. The turnitin.com peer assessment functions primarily focused on generating end comments rather than marginalia. Reviewers were able to tag specific locations in the text that could be used in the end comment to indicate where a particular problem existed; however, this function was not obvious and most students did not use it. In addition, the specific commenting prompts were separate from the ratings prompts, which could allow one to create a reviewing assignment that utilized more fine-grained evaluation dimensions

and broader commenting dimensions. Finally, the reviews were anonymous—that is, a pseudonym was used to identify both the writer and the reviewer.

For this writing task, students were expected to write a three-page paper in which they evaluated whether MSNBC.com, a US digital news provider, accurately reported a psychological study—applying concepts from the Research Methods chapter covered in lecture and lab in the prior week. After the first draft deadline passed, four papers were distributed to participants for review based on the reviewer ability condition they were assigned (i.e., participants were randomly assigned to a reviewer ability condition—either low reviewer or high reviewer—and four papers from peers designated as low or high reviewers were randomly chosen to be the reviewers). As reviewers, participants were blind to the ability level of the authors. Participants were able to access the peer feedback online once the reviewing deadline had passed. The participants were given one week to revise their draft based on the peer feedback. As authors, participants were blind to the ability level of the reviewers. The TAs also provided final grades for the paper.

The dependent variables included the initial and revised text quality, number of high-level criticism comments, problems, and solutions received, number of low level criticism comments received, the number of implemented high-level criticism comments, problems, and solutions, number of implemented low-level criticism comments, the quality of the revisions based on high-level criticism comments, problems, and solutions, and the quality of the revisions based on low-level criticism comments.

### 2.5 Review Support Structures

Participants were provided with a detailed rubric to use for the reviewing task, which shaped what comments were available for analysis in this study. The rubric included commonly-used general reviewing suggestions (e.g., be nice, be constructive, be specific) and specific guidelines, which described the three reviewing dimensions that have been applied in many disciplinary writing settings: flow, argument logic, and insight. For each commenting dimension, a number of questions were provided to prompt the reviewer to consider the paper using several particular lenses. The *flow* dimension focused on whether the main ideas and the transitions between the ideas were clear (e.g., Did the writing flow smoothly so you could follow the main argument? Did you understand what each argument was and the ordering of the points made sense to you?). The *argument logic* dimension focused on whether the main ideas were appropriately supported and whether obvious counter-arguments were considered (e.g., Did the author just make some claims or did the author provide some supporting arguments or evidence for those claims? Did the author consider obvious counter-arguments, or were they just ignored?). The *insight* dimension focused on whether a perspective beyond the assigned texts and other course materials was provided (e.g., Did the author just summarize what everybody in the class would already know from coming to class and doing the assigned readings, or did the author tell you something

new? Did the author provide an original and interesting alternative explanation?). The purpose of these specific guidelines was to direct the participants' attention primarily towards global writing issues (Wallace & Hayes, 1991).

Finally, participants rated the quality of the papers using a 5-point scale (1–'Very Poor' to 5–'Very Good'). They rated six aspects of the paper within the three commenting dimensions of flow (i.e., how well the paper stayed on topic and how well the paper was organized), argument logic (i.e., how persuasively the paper made its case, how well the author explained why causal conclusions cannot be made from correlational studies, and whether all the relevant information from the research article was provided), and insight (i.e., how interesting and original the paper's conclusion was to the reviewer). For each rating, participants were given descriptive anchors to help determine which rating was most appropriate.

## 2.6 Coding Process

To fully investigate the effects of ability, the students' papers, the comments received, and the revisions that were implemented were coded. Documents were rated for quality to validate the ability difference groupings. Peer comments were coded to examine how their substance varied by ability. Finally, the revisions were coded to examine ability effects on both rate of implementation of each comment and quality of revisions.

*Quality of writing.* Two outside writing experts (i.e., rhetoric graduate students with extensive writing teaching experience) rated the quality of the first drafts. The rubric used by the participants was elaborated for the expert coders in order to examine quality at a more fine-grained level (see Appendix A for rubric details). For example, the students' dimension of "how well this paper was organized" was further divided into two dimensions for the experts: "how well this paper was organized around a main idea" and "how well transitions connected paragraphs". A similar 5-point scale (1–'Very Poor' to 5–'Very Good') was used. The final inter-rater reliability was high (ICC = 0.84; Shrout & Fleiss, 1979).

*Coding peer feedback.* The feedback was coded to determine how the amount and type of comments varied as a function of author ability and reviewer ability. Pairs of undergraduate research assistants (RAs) coded each feature and focus (i.e., problems, solutions, low-level, high-level)—Kappa values are presented for each type. Because some of the inter-rater reliabilities were moderate, comments were exhaustively double-coded, and coders were required to come to consensus on all disagreement cases to improve effective reliability and reduce coding noise.

First, all the feedback received for each reviewing dimension for each writer was compiled. Then, when necessary, each piece of feedback on a given reviewing dimension was further segmented into individual comments based on idea units representing unique issues to be addressed because reviewers frequently commented

about multiple issues within one dimension. For example, one comment about prose flow might mention a problem in the transitions, the relative lack of use of examples, and a specific problem in word choice; this one piece of feedback would then be divided into three segments. If multiple sentences explained the nature and location of a problem and then how it might be repaired, this was treated as one idea unit. This segmentation process produced a total of 7,641 comments to be coded and analyzed.

Second, each comment was coded for the presence or absence of two independent features: problem comments described something wrong with the paper (e.g., "The writer did not offer insight into causal and correlational relationships."; Kappa = .85), and solution comments suggested how to fix a problem or improve the quality of the paper (e.g., "Also, I would suggest writing a stronger conclusion to the end of the paper."; Kappa = .90). In addition to being important features of comments that influence how readily authors implement the feedback, comments that contain neither problem nor solution description cannot be further coded for implementation or revision quality (i.e., for praise or summary comments, there is nothing to implement). Next, all comments that were previously coded as either problem or solution (i.e., criticism comments) were coded for the focus (i.e., low-level, high-level—Kappa = .77). Low-level comments focused on issues dealing with the literal text choice—usually at a word level (e.g., "Where you say 'the hypotheses and whether those hypotheses were proven', I think you would say 'that hypothesis' or 'the hypothesis' because it's just one hypothesis."), whereas high-level comments focused on issues with missing, incorrect, or contradictory content or high-level writing issues like clarity, use of transitions, strength of arguments, provision of support and counter-arguments (e.g., "I do not understand what the argument is as it isn't very clear."). Given that lower-ability writers are less likely to address global issues, it is especially important to determine whether certain feedback features (i.e., problem descriptions or suggested solutions) better support their revision behaviors. Figure 2 illustrates the relationship between the feedback provided, segmented comments, and the types of feedback coded. An example of how one piece of feedback was segmented and coded can be found in Appendix B.

*Implementation Rate.* To examine the implementation rate of the comments, the same two writing experts coded whether the writer implemented a revision that addressed the issue identified in each criticism comment (*Kappa* = .74—percent agreement was 89%). Microsoft Word's "compare documents" function applied to the 2nd vs. 1st drafts was used to facilitate this process. As long as the writer appeared to attempt a revision based on the comment, it was coded as implemented (see Appendix C for examples). A small percentage of comments (5%) were excluded from analysis for being too vague or unclear to determine whether they were implemented (e.g., "If anything, the paper should be a little spruced up").

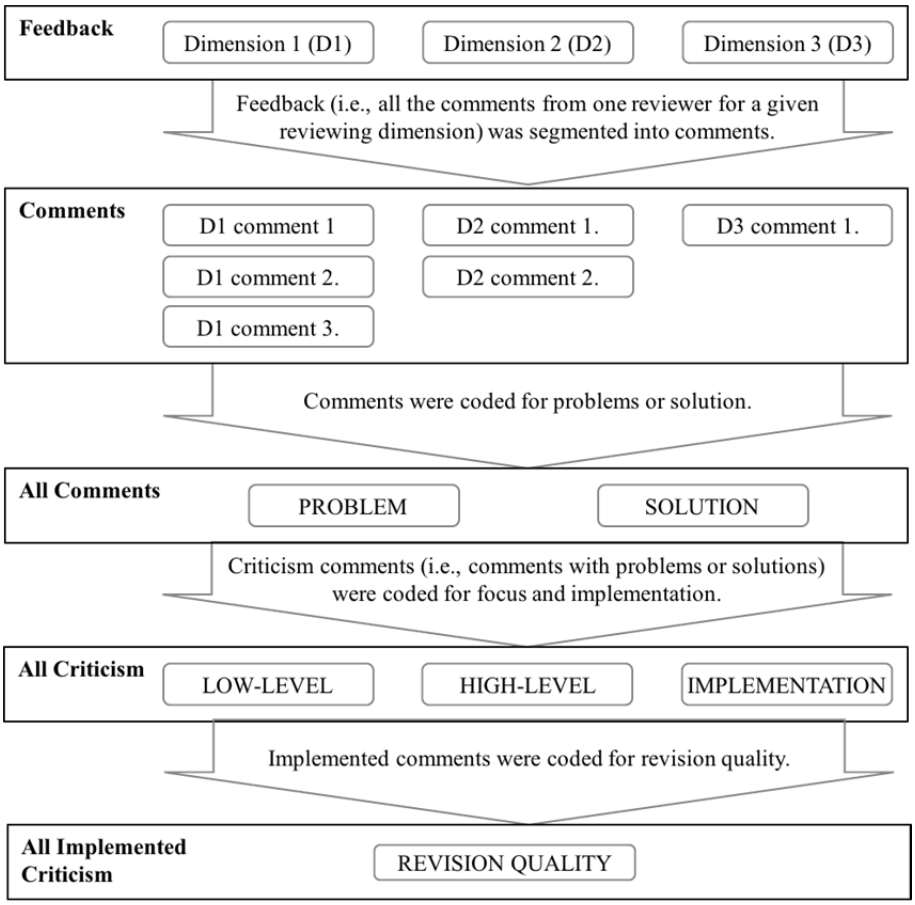



***Figure 2.*** Coding process.

*Revision quality*. For comments that were implemented, the same two writing experts rated the quality of the revision associated with each implemented comment, with high inter-rater reliability (*ICC* = .77). This quality rating was on a binary scale aligned with the original document quality rubric: a rating of 0 indicated either no change in the quality of the paper or a decrease in the quality of the paper (a rare outcome, and hence collapsed with no change); and a rating of 1 indicated an increase in the quality of the paper (see Appendix C for examples).

## 3. Results and Discussion

### 3.1 Overview

To analyze the effects of author ability and reviewer ability on the three products of peer assessment (i.e., initial text, peer feedback, and revisions), each dependent variable (i.e., draft improvement, number of comments, implementation rate, and revision quality) was analyzed using a 2 x 2 between-subjects ANOVA with author ability (i.e., high author vs. low author) and reviewer ability (i.e., high reviewer vs. low reviewer) as between-subjects variables. To uncover the source of the significant interactions, independent t-tests were performed comparing high reviewers to low reviewers for high authors and low authors separately as well as comparing high authors to low authors for high reviewers and low reviewers separately.

**Table 3.** Proportion of Variance Explained at Each Level for the Unconditional HLM Models

| DV | L1 (reviews) | L2 (reviewers) | L2 (authors) |
|---|---|---|---|
| **Peer Feedback** | | | |
| Overall Criticism | 0.57 | 0.31 | 0.12 |
| High Level Criticism | 0.58 | 0.27 | 0.14 |
| High Level Problems | 0.69 | 0.27 | 0.04 |
| High Level Solutions | 0.57 | 0.32 | 0.11 |
| Low Level Criticism | 0.80 | 0.18 | 0.02 |
| **Implementation Rate** | | | |
| Overall Criticism | 0.55 | 0.05 | 0.40 |
| High Level Criticism | 0.51 | 0.07 | 0.43 |
| High Level Problems | 0.54 | 0.05 | 0.41 |
| High Level Solutions | 0.61 | 0.08 | 0.32 |
| Low Level Criticism | 0.69 | 0.07 | 0.24 |
| **Revision Quality** | | | |
| Overall Criticism | 0.63 | 0.06 | 0.31 |
| High Level Criticism | 0.61 | 0.05 | 0.33 |
| High Level Problems | 0.57 | 0.06 | 0.37 |
| High Level Solutions | 0.75 | 0.02 | 0.23 |
| Low Level Criticism | 0.28 | 0.00 | 0.72 |

Additionally, for each of the outcomes (i.e., number of comments, implementation of rate and revision quality), we conducted a series of two-level, cross-classified,

hierarchical regression models (Raudenbush & Bryk, 2002). This type of statistical model takes into account the complex data structure—that is, the dependent variables were not independent observations (i.e., 717 reviews were provided to 189 authors from 369 reviewers), and they were not cleanly nested (i.e., reviews were nested within both authors and reviewers, but reviewers were not nested within authors or vice versa). These analyses revealed that the proportion of variance explained at Level-2 (i.e., author level and reviewer level) depended on the type of dependent variable (see Table 3).

In general, the reviewer controlled the number of comments (i.e., little to no variance was explained at the author level), and the author controlled the number of implemented comments and revision quality (i.e., little to no variance was explained at the reviewer level). We revisit this interesting pattern in the General Discussion. Most importantly for the central research questions tested here, the effects found of the author and reviewer ability using the HLM models were similar to those found using simple ANOVA models. For simplicity, we report the details from the ANOVA results in the sections that follow.

All descriptive and inferential statistics for the number of peer comments, implementation rate, and average revision quality can be found in Table 4. As an indicator of effect size, eta squared (i.e., $\eta^2$—proportion of variance in the dependent variable accounted for by the independent variable(s) while controlling for other possible variables) was included for all ANOVAs. An $\eta^2$ of .01 is considered small, .06 is medium, and .14 is large (Cohen, 1988). Cohen's *d* (i.e., mean difference divided by average standard deviation) was included for all t-tests. Typically, a Cohen's *d* of .3 is considered small, .5 is medium, and .8 is large (Cohen, 1977).

## 3.2 Initial Text

Text quality was rated using a 5-point scale on eight different dimensions. The sum of all eight dimensions (i.e., out of 40 possible points) was analyzed. Because higher-ability writers are able to detect, diagnose, and solve more problems than lower-ability writers (Hayes et al., 1987), high authors were expected to produce higher quality first drafts than low authors. The results supported this prediction. As noted in the Methods section, there was a significant main effect of author ability with the high authors ($M$ = 17.0, $SD$ = 3.8) outperforming the low authors ($M$ = 14.8, $SD$ = 3.2), $F(1,185) = 18.55$, $p < .001$, $d = 0.63$. It is important to note that this document quality effect, though clearly statistically significant, was not large, even though the groups had large writing ability differences on multiple measures. There are a number of factors that generally influence first draft quality (and likely mattered in this context) beyond the author's writing ability, such topical interest, grade goals set for the course, and competing course and work demands (Schiefele, 1999; Svanum & Bigatti, 2006; Wood & Locke, 1987). Thus, it is likely often the case that author ability does not have as large an effect

**Table 4.** Number of Peer Comments, Implementation Rate, and Average Revision Quality by Author Ability and Feedback Source for High-Level and Low-Level Comments

| Reviewer Ability | | Author Ability | | | | | | ANOVA | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Low | | | High | | | Author ability | | Reviewer ability | | Interaction | |
| | | *N* | *M* | *SD* | *N* | *M* | *SD* | *p* | $\eta^2$ | *p* | $\eta^2$ | *p* | $\eta$ |
| HIGH LEVEL | | | | | | | | | | | | | |
| *Peer Comments* | | | | | | | | | | | | | |
| Criticism | L | 47 | 12.4 | 5.9 | 46 | 11.7 | 5.2 | .20 | .01 | .003 | .05 | .62 | .00 |
| | H | 49 | 15.7 | 6.6 | 47 | 14.0 | 7.7 | | | | | | |
| Problems | L | 47 | 7.3 | 4.6 | 46 | 8.0 | 3.9 | .29 | .01 | .01 | .04 | .04 | .02 |
| | H | 49 | 10.7 | 5.4 | 47 | 8.5 | 5.4 | | | | | | |
| Solutions | L | 47 | 7.3 | 4.2 | 46 | 5.9 | 3.6 | .31 | .01 | .16 | .01 | .23 | .01 |
| | H | 49 | 7.4 | 4.6 | 47 | 7.5 | 4.6 | | | | | | |
| *Implementation* | | | | | | | | | | | | | |
| Criticism | L | 506 | 0.38 | 0.49 | 464 | 0.29 | 0.45 | .57 | .00 | .09 | .00 | < | .01 |
| | H | 681 | 0.27 | 0.44 | 577 | 0.33 | 0.47 | | | | | | |
| Problems | L | 301 | 0.37 | 0.48 | 308 | 0.32 | 0.47 | .53 | .00 | .58 | .00 | .01 | .00 |
| | H | 457 | 0.29 | 0.45 | 337 | 0.37 | 0.48 | | | | | | |
| Solutions | L | 301 | 0.36 | 0.48 | 254 | 0.29 | 0.46 | .69 | .00 | .24 | .00 | .03 | .00 |
| | H | 340 | 0.27 | 0.44 | 328 | 0.32 | 0.47 | | | | | | |
| *Revision Quality* | | | | | | | | | | | | | |
| Criticism | L | 189 | 0.30 | 0.51 | 134 | 0.22 | 0.43 | .91 | .00 | .70 | .00 | .02 | .01 |
| | H | 181 | 0.20 | 0.43 | 192 | 0.29 | 0.45 | | | | | | |
| Problems | L | 110 | 0.24 | 0.51 | 99 | 0.18 | 0.41 | .99 | .00 | .44 | .00 | .19 | .00 |
| | H | 131 | 0.21 | 0.43 | 126 | 0.27 | 0.45 | | | | | | |
| Solutions | L | 107 | 0.33 | 0.53 | 73 | 0.21 | 0.44 | .97 | .00 | .27 | .00 | .01 | .02 |
| | H | 91 | 0.15 | 0.39 | 103 | 0.27 | 0.45 | | | | | | |
| LOW-LEVEL | | | | | | | | | | | | | |
| Peer Comments | L | 47 | 3.3 | 2.8 | 46 | 2.3 | 2.0 | .03 | .03 | .40 | .00 | .71 | .00 |
| | H | 49 | 3.4 | 2.9 | 47 | 2.7 | 2.2 | | | | | | |
| Implementation Rate | L | 127 | 0.41 | 0.49 | 88 | 0.36 | 0.48 | .10 | .01 | .54 | .00 | .51 | .00 |
| | H | 152 | 0.47 | 0.50 | 119 | 0.36 | 0.48 | | | | | | |
| Revision Quality | L | 51 | 0.04 | 0.40 | 32 | 0.00 | 0.00 | .72 | .00 | .45 | .00 | .45 | .00 |
| | H | 71 | -0.01 | 0.12 | 42 | 0.00 | 0.22 | | | | | | |

*Note*. For peer comments, N is the number of students, and M is the mean number of comments received. For implementation rate, N is the total number of comments, and M is the proportion of comments that were implemented. For revision quality, N is the number of implemented comments, and M is the proportion of comments that improved the quality of the paper.

on the quality of documents being reviewed as one might have otherwise expected. As expected, because there was no reviewing prior to submission, there was neither a significant main effect of reviewer ability, $F(1,185) < 1$, $p = .65$, nor a significant interaction between author ability and reviewer ability, $F(1,185) < 1$, $p = .86$.

### 3.3 Peer Feedback

The amount of four types of feedback was analyzed. First, criticism comments represent feedback that is actionable. Next, these comments were further categorized as describing problems or offering solutions. These two features represent important revision skills. Note that problems and solutions were coded independently, and thus would not add up to the amount of criticism comments. Finally, these comments were also categorized as focusing on low-level issues or high-level issues. These features may be especially important for supporting lower-ability writers who are less likely to address global issues.

To examine the effect of author ability and reviewer ability on the peer feedback received, we collapsed the data by authors—that is, for each author the number of comments received for each feedback feature was summed. Overall, the total number of all comments an author received did not differ by author ability, $F(1,185) = 2.51$, $p = .12$, nor reviewer ability, $F(1,185) < 1$, $p = .64$, and there was not a significant interaction between author ability and reviewer ability, $F(1,185) < 1$, $p = .47$. On average, students received a total of 40.4 comments ($SD = 9.2$). The lack of an effect on the number of comments was convenient for in-depth analyses of these comments because there is no need to distinguish between absolute frequency and relative frequency. Because low-level feedback may be more readily produced, more easily implemented, and of lower importance than high-level feedback, we analyzed high-level comments separately from low-level comments.

Does the author ability affect the amount and types of peer feedback received? Since high authors produced higher quality texts, they were expected to receive fewer criticism comments than low authors. The results of the current study do not support this prediction. There was no main effect of author ability for the number of high-level criticism comments, $F(1,185) = 1.65$, $p = .20$. On average, students received 13.5 high-level criticism comments ($SD = 6.6$). Moreover, there was no main effect of author ability for the number of high-level problems, $F(1,185) = 1.14$, $p = .29$, or the number of high-level solutions, $F(1,185) = 1.06$, $p = .31$. On average, students received 8.7 ($SD = 5.0$) comments that described high-level problems and 7.1 ($SD = 4.3$) comments that described high-level solutions. These results suggest that in general college students continue to struggle with high-level writing issues regardless of their writing ability.

Does reviewer ability affect the amount and types of peer feedback received? Again, because higher-ability writers are able to detect, diagnose, and solve more problems in their own writing than lower-ability writers (Hayes et al., 1987), all authors were expected to receive a greater number of criticism in general (and solutions and localized comments in particular) from high reviewers than from low reviewers. The

results partially supported this prediction. There was a significant main effect of reviewer ability for the number of high-level criticism comments received, $F(1,185) = 8.92$, $p = .003$. Overall, authors received a greater number of high-level criticism comments from high reviewers than low reviewers. Of these comments, authors received a greater number of high-level problems from high reviewers than low reviewers, $F(1,185) = 7.52$, $p = .01$. There was no main effect of reviewer ability on the number of high-level solutions provided, $F(1,185) = 1.97$, $p = .16$. These results suggest that a student's writing ability could also reflect his or her ability to detect and diagnose problems in peers' texts, but not necessarily his or her ability to suggest revision strategies.

Does reviewer ability also moderate effects of author ability (i.e., initial text quality) on the amount and types of peer feedback received? Since higher-ability writers have more refined reviewing skills than lower-ability writers (Hayes et al., 1987), high reviewers were expected to better detect the differences in high versus low text quality than would low reviewers. This prediction was partially supported. Although there was not a significant interaction for the overall number of criticism comments, $F(1,185) < 1$, $p = .62$, or the number of solutions, $F(1,185) = 1.43$, $p = .23$, there was a significant interaction between author ability and reviewer ability for the number of high-level problems, $F(1,185) = 4.47$, $p = .04$ (see Figure 3A). High reviewers described significantly more high-level problems to low authors than to high authors, $t(94) = 2.05$, $p = .04$, but low reviewers offered a similar number of high-level problems comments to low authors and high authors, $t(91) < 1$, $p = .41$.

*Low-level comments.* Given that the reviewing prompts focused on high-level issues and advised against commenting on low-level problems, students received fewer low-level criticism comments overall. However, as expected, low authors received slightly more low-level criticism comments than high authors, $F(1,185) = 5.09$, $p = .03$. There was not a significant main effect of reviewer ability on low-level criticisms, $F(1,185) < 1$, $p = .40$, nor was there a significant interaction between author ability and reviewer ability, $F(1,185) < 1$, $p = .71$ (see Figure 3B).

### 3.4 Revisions

To examine the effect of author ability and reviewer ability on the revisions, we analyzed revisions in two ways: implementation rate and revision quality. Implementation rate only included implementable comments ($N = 2,714$)—that is, the comments must include a description of a problem or suggest a solution to be included in this analysis. Also note that revision quality could only be coded for comments that were implemented ($N = 892$). Both implementation rate and revision quality were analyzed at the comments level. Again, we analyzed high-level comments separately from low-level comments. As noted earlier, although comments are nested within authors and reviewers, more complex nested models that accounted for these

dependencies among the data produced similar results and therefore only the simpler ANOVA results are described.

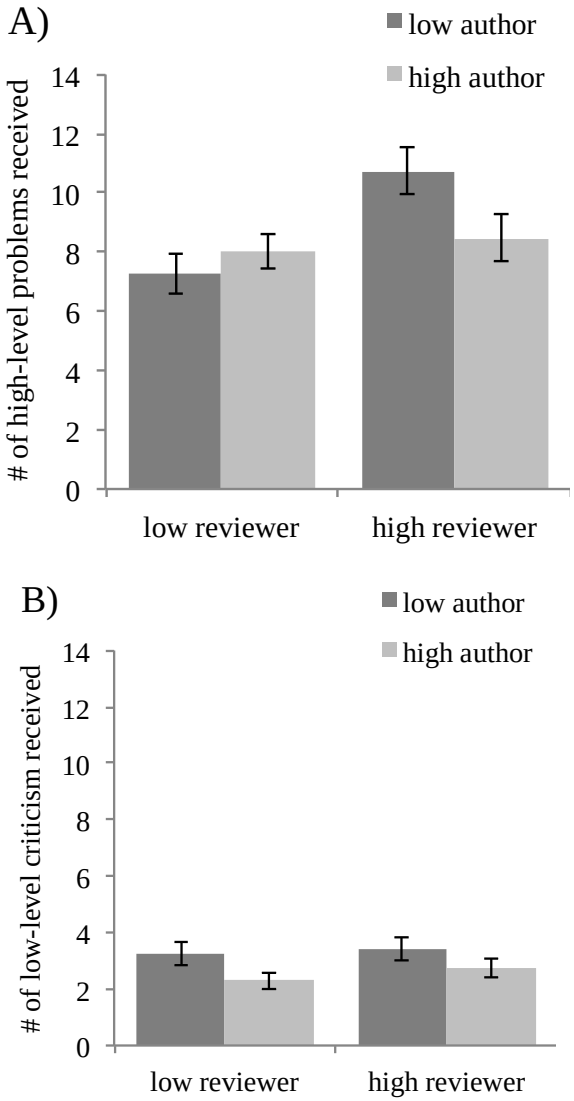



***Figure 3.*** The interaction between author ability and reviewer ability for the number of A) high-level problems received and B) low-level criticism received.

Does author ability affect the amount and quality of revisions made? Not only do higher-ability writers detect, diagnose, and solve more problems in their own writing than lower-ability writers, they also choose more effective revision strategies (Hayes et al., 1987). Therefore, high authors were expected to make more revisions that were of higher quality than low authors. These predictions about author ability were not supported in this context of peer feedback. There was not a significant main effect of author ability for the implementation rate, $F(1,2224) < 1$, $p = .57$, or revision quality, $F(1,692) < 1$, $p = .91$, of high-level criticism comments. As is commonly the case in multi-peer assessment, students received many comments but were quite selective in which comments were addressed—on average, students implemented only 32% of the high-level criticism comments ($SD = 0.47$). Turning to more specific categories of feedback, there were no main effects of author ability for the number of implemented problems, $F(1,1399) < 1$, $p = .53$, or solutions, $F(1,1219) < 1$, $p = .69$.

When students implemented these high-level criticism comments, the average revision quality was 0.25 ($SD = 0.46$)—that is, approximately one quarter of revisions led to a significant improvement in document quality. Mirroring the implementation rate pattern, there were no main effects of author ability for the revision quality of high-level problems, $F(1,462) < 1$, $p = .99$, or solutions, $F(1,370) < 1$, $p = .97$ . These results indicate that, on average, the rates of critical revisions to high-level issues were not driven by author ability.

Does reviewer ability affect the amount and quality of revisions made? Because authors were expected to receive more articulate criticism comments from high reviewers than low reviewers, they were expected to be more likely to make revisions after receiving feedback from high reviewers than low reviewers. Similarly, because high reviewers are likely better able to choose effective solutions (Hayes et al., 1987), the revisions made from implementing high reviewers' comments were expected to be of higher quality. These simple main effect predictions were not supported: there were no significant main effects of reviewer ability for the implementation rate, $F(1,2224) = 2.87$, $p = .09$, or revision quality, $F(1,692) < 1$, $p = .70$.

Does author ability moderate the effect of reviewer ability on the amount and quality of revisions made? Higher-ability writers have a more sophisticated means-ends repertory (Hayes et al., 1987), which would allow them to address relevant comments even if they lack specific details about the problem or solution. Therefore, high authors were expected to equally address issues detected by both high reviewers and low reviewers despite the likely difference in quality of comments. By contrast, lower-ability writers have fewer solutions available to them and are less able to choose effective solutions (Hayes et al., 1987). Therefore, it was unclear whether low authors would be able to distinguish between more or less effective solutions, successfully implement all effective solutions if unfamiliar strategies were included, or come up with a solution when a problem was described without one. Indeed, there were several interesting interactions.

First, there were significant interactions between author ability and reviewer ability for the implementation rate of high-level criticism comments, $F(1,2224) = 15.13$, $p < .001$, high-level problems, $F(1,1399) = 6.85$, $p = .01$, and high-level solutions, $F(1,1219) = 4.60$, $p = .03$ (see Figure 4A). High authors implemented a similar number of high-level criticism comments, $t(1039) < 1$, $p = .13$, high-level problems, $t(643) = 1.40$, $p = .16$, and high-level solutions, $t(580) < 1$, $p = .51$, regardless of reviewer ability. However, low authors implemented more high-level criticism comments, $t(1185) = 4.08$, $p < .001$, high-level problems, $t(756) = 2.34$, $p = .02$, and high-level solutions, $t(639) = 2.41$, $p = .02$, from low reviewers than from high reviewers.

Moreover, there were significant interactions between author ability and reviewer ability for the revision quality of high-level criticism comments, $F(1,692) = 5.65$, $p = .02$, and high-level solutions, $F(1,370) = 6.25$, $p = .01$ (see Figure 4B). Similarly, high authors' revisions were of similar quality for high-level criticism comments, $t(324) = 1.4$, $p = .16$, and high-level solutions, $t(174) < 1$, $p = .33$, regardless of reviewer ability. By contrast, low authors' revisions were of higher quality when high-level criticism comments, $t(368) = 1.98$, $p = .05$, and high-level solutions, $t(196) = 2.58$, $p = .01$, were received from low reviewers than from high reviewers. The interaction between author ability and reviewer ability for the revision quality of specifically problems was not statistically significant, $F(1,462) = 1.74$, $p = .19$. These results indicate that high authors are equally capable of improving their paper with feedback from either a low reviewer or high reviewer, however, low authors benefit the most from feedback provided by low reviewers.

*Low-level comments*. In contrast to the patterns for criticism of high-level problems, the low-level feedback did not show any effects of author ability or reviewer ability, nor was there a significant interaction between author ability and reviewer ability for implementation rate or quality of implementations. Specifically, there was not a significant main effect of author ability on the implementation rate of low-level criticisms, $F(1,482) = 2.78$, $p = .10$, nor was there a main effect of reviewer ability, $F(1,482) < 1$, $p = .54$. Furthermore, there was not a significant interaction between author ability and reviewer ability, $F(1,482) < 1$, $p = .51$. On average, students implemented 41% of the low-level criticism comments ($SD = .49$), a higher rate than the implementation of high-level criticism.

On average, revisions to low-level issues rarely produced significant improvements in the draft ($M = .01$, $SD = .24$), as one would expect from low level issues. Thus, it is not surprising that there was no main effect of author ability on the revision quality of low-level criticisms, $F(1,192) < 1$, $p = .72$, no main effect of reviewer ability, $F(1,192) < 1$, $p = .45$, nor an interaction between author ability and reviewer ability, $F(1,192) < 1$, $p = .45$.

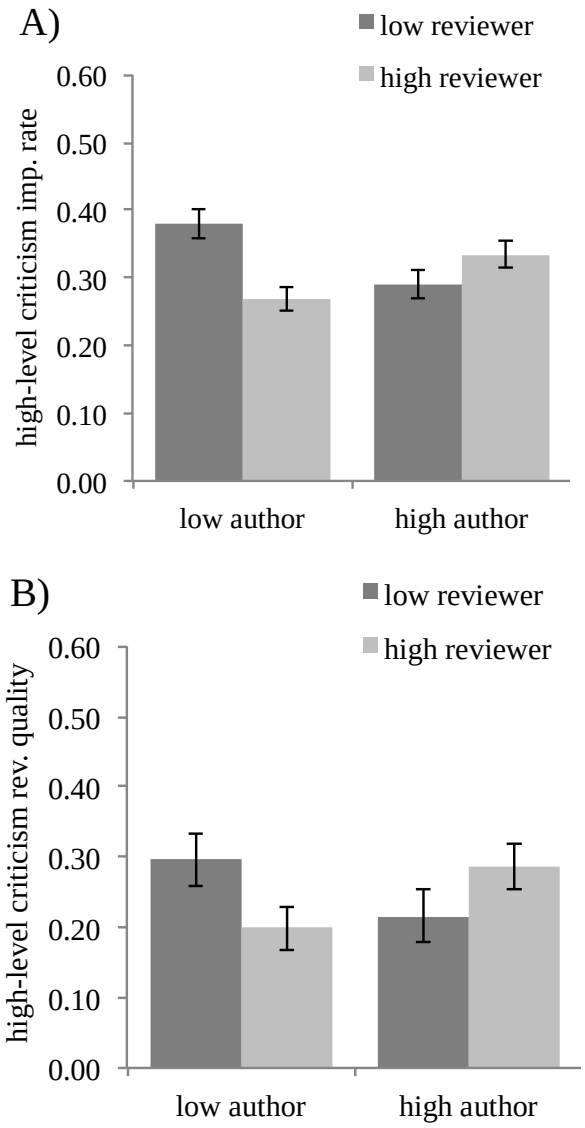



***Figure 4.*** The interaction between author ability and reviewer ability for A) the implementation rate of high-level criticism comments and B) the revision quality of high-level criticism comments.

## 4. General Discussion

### 4.1 Summary of Results

A significant amount of research has focused on how student ability affects the benefits of peer tutoring (Cohen, Kulik, & Kulik, 1982) and cooperative learning (Lou et al., 1996). More recent research has focused on peer-assisted learning of writing, including collaborative writing (Van Steendam, Rijlaarsdam, Van den Bergh, & Sercu, 2014) and peer assessment of writing (Patchan et al., 2013). The goal of the current study was to fully document the relationship between author ability and reviewer ability and its effect on peer assessment from initial text production through revision. We also developed a new peer assessment model that includes the effects of student ability based on the Flower and Hayes writing model (see Figure 1) to help organize and predict a complex set of interacting effects at multiple points along the writing and reviewing process.

We begin with a brief empirical summary of the observed effects, which comprised a mix of expected effects but also theoretically and practically important surprises. First, the almost definitional predictions about the initial text were supported: High authors produced higher quality texts than low authors. Second, two of the three predictions about peer feedback were partially supported. Although there were no significant author ability effects, there were several significant reviewer ability effects as well as interactions between author ability and reviewer ability. High reviewers provided more criticism comments than low reviewers. They also described more problems than low reviewers, but this main effect was likely driven by the interaction with author ability—that is, compared to low reviewers, high reviewers were more likely to describe problems in low authors' texts than high authors' texts. Finally, two of the six predictions about revisions were supported. Very interestingly, there were no author ability or reviewer ability effects on implementation or revision quality. Instead, there were many significant interactions between author ability and reviewer ability. In general, high authors were equally likely to implement and improve their paper using feedback from low reviewers and high reviewers. Low authors, however, were more likely to implement and improve their paper using feedback from other low reviewers rather than from high reviewers. These results support the general framework with the interactions frequently dominating main effects. Particular effects are considered in more depth in the next section.

### 4.2 Theoretical Implications of Observed Empirical Patterns

Hayes et al. (1987) observed that higher-ability writers were able to detect, diagnose, and solve more problems in their own writing than lower-ability writers. Higher-ability writers also have more refined reviewing skills than lower-ability writers. They choose more effective solutions and have a more sophisticated means-ends repertory. However, it is possible that the cognitive abilities driving reviewing-while-writing

processes and reviewing-others'-texts processes may differ. Higher-ability writers may employ tacit knowledge that guides the production of successful texts, but they may lack a similar level of cognitive or meta-cognitive skill in identifying, describing, or offering solutions to problems in others' texts. Here we tested and observed effects of these writing ability differences on several aspects of the peer assessment process, focusing the perspective of what feedback authors received and how they revised.

Most obviously, as authors, students' writing ability affected the quality of the initial draft—that is, high authors produced higher quality initial drafts than low authors. This quality difference was likely at least partially a result of high authors detecting, diagnosing, and solving more problems during the process of writing the first draft. Interestingly, large ability differences had a small effect on document quality. Thus, a theory of peer assessment effects that (sensibly) includes document quality factors will need to include other individual difference factors than author ability, perhaps such as writer motivation or more localized knowledge of the writing topic (Hayes, 1996; Pajares, 2003).

More interesting theoretically, as reviewers, students' writing ability affected the amount and types of peer feedback. In general, reviewer ability had a larger effect than author ability on the content of the feedback. Similar to the Patchan et al. (2013) study, high reviewers offered more criticism comments that specifically described problems. Overall, being able to detect, diagnose, and solve more problems, as the model predicted, appeared to help high reviewers provide more feedback to their peers. However, lower quality papers (i.e., papers written by low authors) did not receive more criticism than higher quality papers. Thus models of peer assessment should not assume that reviewers will comment on all detected problems, but rather there is likely some overall amount of feedback that a given reviewer will produce. Indeed, the cross-nested models of the data revealed that the reviewers accounted for far more variance in the amount of feedback produced than the author/document itself (especially for high level issues).

Yet there were also important interactions. Reviewer ability moderated the effect of author ability on the number of problems described. Again, similar to the Patchan et al. (2013) study, high reviewers described more problems to low authors than high authors. These results could reflect differences in number of problems (i.e., papers by low authors have more problems) crossed with differences in ability to detect problems (i.e., high ability reviewers are better able to detect high-level problems), suggesting that high reviewers, who have more refined reviewing skills, may better notice differences between the higher quality texts and lower quality texts. The pattern may also reflect students' commenting style rather than how many problems they are able to detect, diagnose, and solve. For example, as part of students' task definition for providing feedback, they may have a limit on the number of problems that need to be detected with a higher limit for high reviewers than low reviewers.

Finally, not previously examined in the Patchan et al. (2013) study, author ability moderated the effect of reviewer ability on both the amount and quality of revisions. In

general, characteristics of the author accounted for much of the variability in amount and quality of revisions. Specifically, low authors were more likely to implement comments from low reviewers at a higher level of quality than from high reviewers, whereas high authors benefitted equally from high reviewers and low reviewers. Although high reviewers detected and described more high-level problems, they may have done so at a level that was not helpful to the low authors. Low reviewers, who comment on fewer high-level issues, may communicate in a way that helps other low authors better. By contrast, high authors, who have a more sophisticated means-ends repertory, were more capable of handling any issue brought up by high reviewers or low reviewers. Thus, a model of peer assessment needs some consideration of a zone of proximal development, where comments at the writer's ability level are needed to produce useful revision.

There were two effects of author ability and reviewer ability that were surprisingly missing. First, author ability (i.e., via initial text quality) did not affect the amount or types of peer feedback received. High authors and low authors received similar number of criticism comments despite the high authors having higher quality texts. This lack of a difference could be a result the relatively small initial differences in text quality. Alternatively, reviewers may have a fixed level of effort or amount of feedback they are willing produce, regardless of the number of issues they could detect. Indeed reviewer variability did account for a significant amount of variation in amount of feedback produced, ruling out the possibility that the scaffolded nature of the reviewing situation (i.e., a fixed number of prompts) led students to always produce the same number of comments.

Second, while other characteristics of authors appeared to matter, neither author ability nor reviewer ability per se directly affected the amount and quality of revisions. Both high authors and low authors implemented a similar number of comments at a similar level of quality. Likewise, the same proportion of comments from high reviewers and low reviewers were implemented. These results suggest that even though high authors produced higher quality initial drafts, all drafts had much to be improved, which was evident in their initial draft scores (i.e., less than 50% of the possible points). Perhaps students vary in their thresholds for the amount of revising they are willing to do, but ability is not the predictor of this threshold. In a related way, multi-peer assessment may consistently produce feedback that is above the threshold for most students. That is, given that so much improvement was needed and so many comments were provided, students might have received more comments than they were willing to address—on average, they only implemented a third of the comments received. Consequently, they were selective in which comments to implement and might have chosen comments that were easier to implement rather than making a distinction between effective comments and ineffective comments—on average, two-thirds of the comments authors implemented did not result in a significant improvement in the draft.

Overall, these results not only suggest that writing ability affects how well one writes (i.e., initial draft quality), but also writing ability affects one's ability to review

peers' texts (i.e., how much and the type of peer feedback). Theories of learning from peer assessment must take into account both reviewer ability and interactions between author ability and reviewer ability. Specifically, theories must explain why students appear to benefit more from being grouped with peers of similar writing ability. Although higher-ability students will likely improve their texts using any peers' comments, lower-ability students appear to benefit more from receiving feedback from other lower-ability peers who can offer feedback at their own level. In general, models of writing revision that include revision from external feedback should focus less on general abilities to detect and diagnose problems and focus more on student's abilities to implement provided suggestions.

### 4.3 Caveats and Future Directions

There are a few caveats to these findings that must be considered. First, several methodological decisions could have affected the power of this study. Given the instructional context of the current study, all students' texts needed to be reviewed regardless of their quality. Furthermore, students needed to be assigned peers' papers to review shortly after the deadline for the writing assignment. In order to accommodate these pragmatic issues, as well as for future instructional applications, we utilized an indirect measure of writing ability as a proxy for reviewer ability and text quality. In addition, we categorized students as high reviewers and low reviewers and authors as low and high ability by using a median split of the writing ability measure. Although we believe that a lower powered study was a reasonable tradeoff for higher external validity, future research should test the effects of writer ability with more direct measures of writing ability and with larger ability differences. For research purposes, direct measures of reviewer ability and text quality could also be examined to more directly examine the effects of those variables. Future research should also examine additional levels of ability (e.g., high, medium, low). Aligned with Vygostsky's zone of proximal development, prior research has demonstrated that optimal groupings may not be homogeneous nor with distal heterogeneous peers (i.e., high-ability peers and low-ability peers), but rather students benefited the most by working with peers in adjacent levels—that is, low-ability peers with medium-ability peers and medium-ability peer with high-ability peers (Webb, 1989).

Second, the current study focused on the form of the feedback rather than the quality according to expert ratings. One reason for not measuring the quality of the comments was because the quality is relative to author. Although an expert may think a comment is of high quality, if the comment is not at the level of the student and he or she is not able to successfully revise the document based on the information provided, the comment is not very effective. However, it is likely that students do not consistently select effective peer comments. Therefore, future research should more closely examine various dimensions of quality of peer feedback and the selection process students use to choose which comments to implement.

Finally, future research should examine what other factors affect students' revision process. Consistently, students choose to only implement about a third of the peer feedback that they receive. There are many possible reasons for this decision: more affective elements of the reviews (e.g., harshness of criticism, inclusion of mitigating language) may have been influential, most of their peer feedback may be ineffective and should not be implemented, most of their peer feedback may not be written at their level causing them to not understand the problem, they may receive too much feedback and experience an overload of information, they may not have much motivation for making revisions to their text, or they may not have developed appropriate strategies for revising with feedback. Future studies should systematically examine how students' motivation and knowledge about revision influences which comments they implement. Finally, future studies need to examine the cumulative impact of peer feedback from different sources on growth in writing ability, rather than just on revision behavior.

### 4.4 Practical Implications

Based on the findings from the current study, lower-ability students may benefit more from being grouped with other lower-ability students, while higher-ability students may benefit equally from being grouped with either lower-ability students or higher-ability students. Pragmatically, these results imply that students grouped with peers of similar ability could benefit all students. Higher-ability students will likely revise their texts successfully regardless of who there are partnered with, but the lower-ability students may need feedback at their own level. Other lower-ability peers seem to be better equipped to provide this type of feedback. At the same time, it is also important to note that both the main effects of reviewer ability and the interaction effects were of relatively small effect sizes, suggesting that students are not as harmed as they often think they are by receiving feedback from lower ability peers, and that peer assessment processes that randomly assign reviewers to papers are not highly problematic.

## Notes

1. The SAT (Scholastic Assessment Test) is a standardized test used for college admissions in the United States. It consists of three sections: the verbal section tests critical reading skills, the writing section tests problem detection skills and grammar and usage knowledge, and the mathematics section tests arithmetic operation, algebra, geometry, statistics, and probability knowledge.
2. Universities in the U.S. typically require a first year composition course, and the university in the present study requires two semesters of composition.

## References


Allal, L., Chanquoy, L., & Largy, P. (Eds.). (2004). *Revision: Cognitive and instructional processes*. Norwell, MA: Kluwer Academic. doi:10.1007/978-94-007-1048-1

Arum, R., & Roksa, J. (2010). *Academically adrift: Limited learning on college campuses*. Chicago: University of Chicago Press. doi:10.7208/chicago/9780226028576.001.0001

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.

Bouzidi, L. h., & Jaillet, A. (2009). Can online peer assessment be trusted? *Educational Technology & Society, 12*(4), 257-268.

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

Chang, C.-F. (2012). Peer review via three modes in an EFL writing course. *Computers and Composition, 29(1),* 63-78. doi:10.1016/j.compcom.2012.01.001

Charney, D. H., & Carlson, R. A. (1995). Learning to Write in a Genre - What Student Writers Take from Model Texts. *Research in the Teaching of English, 29(1)*, 88-125.

Cho, K., & MacArthur, C. (2011). Learning by reviewing. *Journal of Educational Psychology, 103*(1), 73-84. doi:10.1037/a0021950

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. doi:10.1016/j.compedu.2005.02.004

Cho, K., Schunn, C. D., & Charney, D. (2006). Commenting on writing - Typology and perceived helpfulness of comments from novice peer reviewers and subject matter experts. *Written Communication, 23(3)*, 260-294. doi:10.1177/0741088306289261

Cho, K., Schunn, C. D., & Wilson, R. W. (2006). Validity and reliability of scaffolded peer assessment of writing from instructor and student perspectives. *Journal of Educational Psychology, 98*(4), 891-901. doi:10.1037/0022-0663.98.4.891

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. New York, NY: Academic Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

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. doi:10.3102/00028312019002237

Crespo Garcia, R., & Pardo, A. (2010, June). A supporting system for adaptive peer review based on learners' profiles. Paper presented at the Workshop on Computer Supported Peer Review in Education, Pittsburgh, PA.

Crespo, R. M., Pardo, A., & Delgado Kloos, C. (2004). An adaptive strategy for peer review. Paper presented at the 34th ASEE/IEEE Frontiers in Education Conference, Savannah, GA USA. doi: 10.1109/fie.2004.1408636

de Smet, M. J. R., Brand-Gruwel, S., Broekkamp, H., & Kirschner, P. A. (2012). Write between the lines: Electronic outlining and the organization of text ideas. *Computers in Human Behavior, 28*(6), 2107-2116. doi:10.1016/j.chb.2012.06.015

DiPardo, A., & Freedman, S. W. (1988). Peer response groups in the writing classroom: Theoretic foundations and new directions. *Review of Educational Research, 58(2)*, 119-149. doi:10.3102/00346543058002119

Elbow, P. (1973). *Writing without teachers*. New York, NY: Oxford University Press.

Falchikov, N., & Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of Educational Research, 70*(3), 287-322. doi:10.3102/00346543070003287

Fitzgerald, J. (1987). Research on revision in writing. *Review of Educational Research, 57*(4), 481-506. doi:10.3102/00346543057004481

Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. *College Composition and Communication, 32*(4), 365-387. doi:10.2307/356600

Gehringer, E. F. (2001). *Assignment and quality control of peer reviewers.* Paper presented at the American Society for Engineering Annual Conference and Exposition.

Giannoukos, I., Lykourentzou, I., Mpardis, G., Nikolopoulos, V., Loumos, V., & Kayafasa, E. (2010). An adaptive mechanism for author-reviewer matching in online peer assessment. In M. Wallace, I. E. Anagnostopoulos, P. Mylonas & M. Bielikova (Eds.), *Semantics in adaptive and personalized services: Methods, tools, and applications*(pp. 109-126). Berlin Heidelberg: Springer-Verlag. doi: 10.1007/978-3-642-11684-1_7

Gielen, S., Tops, L., Dochy, F., Onghena, P., & Smeets, S. (2010). A comparative study of peer and teacher feedback and of various peer feedback forms in a secondary school writing curriculum. *British Educational Research Journal, 36*(1), 143-162. doi:10.1080/01411920902894070

Hartberg, Y., Gunersel, A. B., Simspon, N. J., & Balester, V. (2008). Development of student writing in biochemistry using calibrated peer review. *Journal of the Scholarship of Teaching and Learning, 2*(1), 29-44.

Hayes, J. R. (1996). A new framework for understanding cognition and affect in writing. In L. W. Gregg & E. R. Steinberg (Eds.), *The Science of writing* (pp. 1-27). Mahwah, New Jersey: Erlbaum.

Hayes, J. R. (2012). Modeling and remodeling writing. *Written Communication, 29(3),* 369-388. doi:10.1177/0741088312451260

Hayes, J. R., Flower, L., Schriver, K. A., Stratman, J. F., & Carey, L. (1987). Cognitive processes in revision. In S. Rosenberg (Ed.), *Advances in applied psycholinguistics Volume 2 Reading, writing, and language learning* (pp. 176-240). New York: Cambridge University Press.

Hsiao, I.-H., & Brusilovsky, P. (2008). Modeling peer review in example annotation. Paper presented at the 16th Annual International Conference on Computers in Education, Taipei, Taiwan.

Kaufman, J., & Schunn, C. (2011). Students' perceptions about peer assessment for writing: Their origin and impact on revision work. *Instructional Science, 39*(3), 387-406. doi:10.1007/s11251-010-9133-6

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. doi:10.3102/00346543066004423

Lundstrom, K., & Baker, W. (2009). To give is better than to receive: The benefits of peer review to the reviewer's own writing. *Journal of Second Language Writing, 18(1)*, 30-43. doi:10.1016/j.jslw.2008.06.002

Moffett, J. (1968). *Teaching the universe of discourse*. Boston, MA: Houghton Mifflin Company.

Mumford, M. D. (1983). Social comparison theory and the evaluation of peer evaluations: A review and some applied implications. *Personnel Psychology, 36(4)*, 867-881. doi:10.1111/j.1744-6570.1983.tb00516.x

National Center for Education Statistics. (2012). The nation's report card: Writing 2011. Retrieved from website: http://nces.ed.gov/nationsreportcard/pdf/main2011/2012470.pdf

Ong, J. (2014). How do planning time and task conditions affect metacognitive processes of L2 writers? *Journal of Second Language Writing, 23*, 17-30. doi:10.1016/j.jslw.2013.10.002

Pajares, F. (2003). Self-efficacy beliefs, motivation, and achievement in writing: A review of the literature. *Reading & Writing Quarterly, 19*(2), 139-158. doi:10.1080/10573560390143085

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

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. doi:10.17239/jowr-2009.01.02.2

Patchan, M. M., Hawk, B., Stevens, C. A., & Schunn, C. D. (2013). The effects of skill diversity on commenting and revisions. *Instructional Science, 41*(2), 381-405. doi:10.1007/s11251-012-9236-3

Patchan, M. M., & Schunn, C. D. (2015). Understanding the benefits of providing peer feedback: How students respond to peers' texts of varying quality. *Instructional Science, 43(5),* 591-614. doi:10.1007/s11251-015-9353-x

Patchan, M. M., Schunn, C. D., & Clark, R. J. (2011). Writing in natural sciences: Understanding the effects of different types of reviewers on the writing process. *Journal of Writing Research, 2(3)*, 365-393. doi:10.17239/jowr-2011.02.03.4

Patchan, M. M., Schunn, C. D., & Correnti, R. J. (in press). The nature of feedback: How feedback features affect students' implementation rate and quality of revisions. *Journal of Educational Psychology*.

Pifarré, M., & Fisher, R. (2011). Breaking up the writing process: How wikis can support understanding the composition and revision strategies of young writers. *Language and Education, 25*(5), 451-466. doi:10.1080/09500782.2011.585240

Raudenbush, S. W., & Bryk, A. S. (2002). Models for cross-classified random effects *Hierarchical linear models: Applications and data analysis methods, second edition* (pp. 373-398). Thousand Oaks, CA: Sage Publications.

Scardamalia, M., & Bereiter, C. (1983). The development of evaluative, diagnostic, and remedial capabilities in children's composing. In M. Martlew (Ed.), *The psychology of written language: Developmental and educational perspectives* (pp. 6-95). New York, NY: Wiley and Sons.

Scardamalia, M., & Bereiter, C. (1987). Knowledge telling and knowledge transformation in written composition. In S. Rosenberg (Ed.), *Advances in applied psycholinguistics: Vol. 2 - Reading, writing, and langauge learning* (pp. 142-175). Cambridge, MA: Cambridge University Press.

Scardamalia, M., Bereiter, C., & Steinbach, R. (1984). Teachability of reflective processes in written composition. *Cognitive Science, 8(2)*, 173-190. doi:10.1207/s15516709cog0802_4

Schiefele, U. (1999). Interest and learning from text. *Scientific Studies of Reading, 3*(3), 257-279. doi:10.1207/s1532799xssr0303_4

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420-428. doi:10.1037/0033-2909.86.2.420

Sommers, N. (1980). Revision strategies of student writers and experienced adult writers. *College Composition and Communication, 31*(4), 378-388. doi:10.2307/356588

Svanum, S., & Bigatti, S. M. (2006). The influences of course effort and outside activities on grades in a college course. *Journal of College Student Development, 47*(5), 564-576. doi:10.1353/csd.2006.0063

Topping, K. J. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research, 68*(3), 249-276. doi:10.3102/00346543068003249

Topping, K. J. (2005). Trends in peer learning. *Educational Psychology, 25*(6), 631-645. doi:10.1080/01443410500345172

Topping, K. J., Dehkinet, R., Blanch, S., Corcelles, M., & Duran, D. (2013). Paradoxical effects of feedback in international online reciprocal peer tutoring. *Computers & Education, 61*, 225-231. doi:10.1016/j.compedu.2012.10.002

Topping, K. J., & Fisher, A. M. (2003). Computerised formative assessment of reading comprehension: field trials in the UK. *Journal of Research in Reading, 26*(3), 267-279. doi:10.1111/1467-9817.00202

Tseng, S.-C., & Tsai, C.-C. (2007). On-line peer assessment and the role of the peer feedback: A study of high school computer course. *Computers & Education, 49*, 1161-1174. doi:10.1016/j.compedu.2006.01.007

Van Steendam, E., Rijlaarsdam, G. C. W., Van den Bergh, H. H., & Sercu, L. (2014). The mediating effect of instruction on pair composition in L2 revision and writing. *Instructional Science, 42*(6), 905-927. doi:10.1007/s11251-014-9318-5

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

Webb, N. M. (1989). Peer interaction and learning in small groups. *International Journal of Educational Research, 13(1)*, 21-39. doi:10.1016/0883-0355(89)90014-1

Wood, R. E., & Locke, E. A. (1987). The relation of self-efficacy and grade foals to academic performance. *Educational and Psychological Measurement, 47*(4), 1013-1024. doi:10.1177/0013164487474017

## Appendix A

Experts' Writing Quality Rubric
FLOW
…how well the paragraphs were developed.

| | |
|---|---|
| 5 – Very Good | All paragraphs stated a point and developed it |
| 4 – Good | Most paragraphs stated a point and developed it |
| 3 – Fair | Some paragraphs stated a point and developed it. All paragraphs introduced a topic, but may not state an explicit point |
| 2 – Poor | Some paragraphs stated a point OR introduced a topic, but did not develop it |
| 1 – Unsatisfactory | No paragraphs stated a point and/or paragraphs shifted topics frequently. |

…how well transitions connected paragraphs.

| | |
|---|---|
| 5 – Very Good | Strong transitions between all paragraphs |
| 4 – Good | Strong transitions between most paragraphs |
| 3 – Fair | Transitions between most paragraphs, but some were weak |
| 2 – Poor | Weak transitions between some of the paragraphs |
| 1 – Unsatisfactory | No transitions between paragraphs |

…how well this paper was organized around a main idea.

| | |
|---|---|
| 5 – Very Good | All paragraphs were connected to the main point |
| 4 – Good | Most paragraphs were connected to the main point |
| 3 – Fair | Some paragraphs were connected to the main point |
| 2 – Poor | Most paragraphs were not connected to the main point |
| 1 – Unsatisfactory | No main point explicitly stated |

ARGUMENT LOGIC
…how well the author evaluated the MSNBC article.

| | |
|---|---|
| 5 – Very Good | All points were supported by concrete evidence or examples |
| 4 – Good | Most points were supported by concrete evidence or examples |
| 3 – Fair | Some points were supported by concrete evidence or examples |
| 2 – Poor | Few points were supported by concrete evidence or examples |
| 1 – Unsatisfactory | No support was provided |

…how well the author explained causal conclusions.

| | |
|---|---|
| 5 – Very Good | Provided a complete and clear explanation (i.e., A • B; B • A; C • [A • B]) |
| 4 – Good | Provided a complete and somewhat clear explanation |
| 3 – Fair | Provided complete but unclear explanation |
| 2 – Poor | Provided an incomplete explanation |
| 1 – Unsatisfactory | No explanation was provided |

…how well the author explained an alternative possibility.

| | |
|---|---|
| 5 – Very Good | Provided an appropriate and clear alternative |
| 4 – Good | Provided an appropriate and somewhat clear alternative |
| 3 – Fair | Provided an appropriate alternative, but did not explain it |
| 2 – Poor | Provided an inappropriate alternative |

| | |
|---|---|
| 1 – Unsatisfactory | No alternative possibility was provided |

…whether all the required information from the research article was accurately provided.

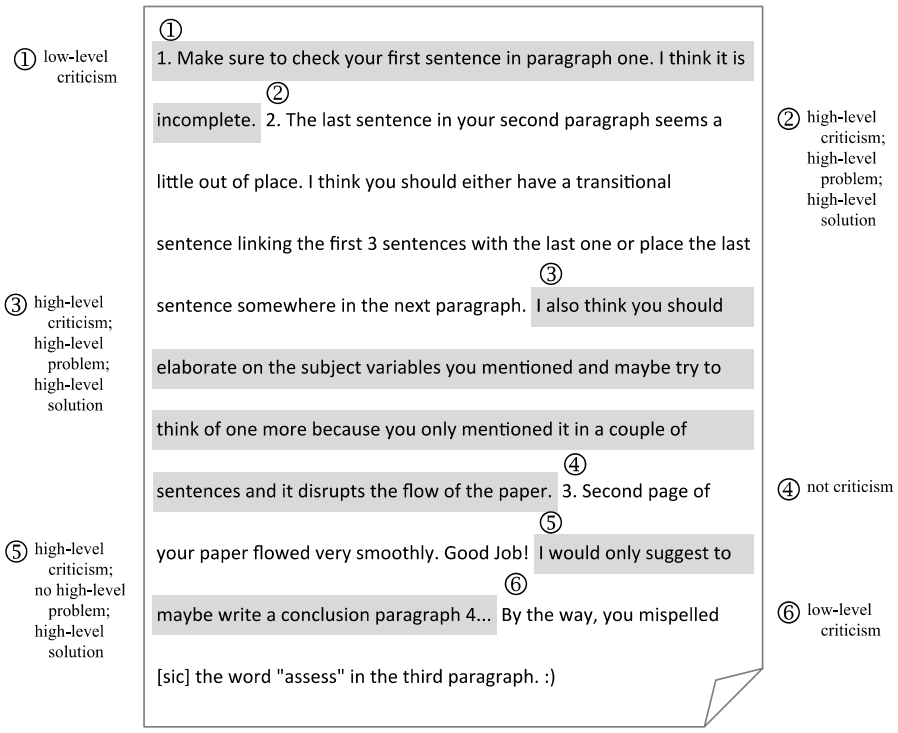
| | |
|---|---|
| 5 – Very Good | The summary accurately included all of the required information |
| 4 – Good | The summary accurately included most of the required information |
| 3 – Fair | The summary accurately included some required information |
| 2 – Poor | The summary included little required information OR the information was inaccurate |
| 1 – Unsatisfactory | No summary of the article |

INSIGHT

…how well the main point was connected to a larger issue.

| | |
|---|---|
| 5 – Very Good | Main point was fully connected to a relevant larger issue throughout the whole paper |
| 4 – Good | Main point was connected to a relevant larger issue |
| 3 – Fair | Some points demonstrated an innovative analysis, but these points were not connected to a relevant larger issue |
| 2 – Poor | One point demonstrated an innovative analysis, but this point was not connected to a relevant larger issue |
| 1 – Unsatisfactory | No points demonstrated an innovative analysis |

## Appendix B


① low-level criticism
1. Make sure to check your first sentence in paragraph one. I think it is
② high-level criticism; high-level problem; high-level solution
incomplete. 2. The last sentence in your second paragraph seems a little out of place. I think you should either have a transitional sentence linking the first 3 sentences with the last one or place the last
③ high-level criticism; high-level problem; high-level solution
sentence somewhere in the next paragraph. I also think you should elaborate on the subject variables you mentioned and maybe try to think of one more because you only mentioned it in a couple of
④ not criticism
sentences and it disrupts the flow of the paper. 3. Second page of
⑤ high-level criticism; no high-level problem; high-level solution
your paper flowed very smoothly. Good Job! I would only suggest to
⑥ low-level criticism
maybe write a conclusion paragraph 4... By the way, you mispelled [sic] the word "assess" in the third paragraph. :)

## Appendix C

Implementation & Revision Quality Coding Scheme

| Rating | Comment | Draft 1 text | Draft 2 text |
|---|---|---|---|
| -1 | Consider revising your introduction The opening quote used, in my opinion shouldn't be the first sentence | According to the original report, the researchers hypothesize that "adults who play video games, compared to nonplayers, would evidence poorer perceptions of their health, greater reliance on Internet-facilitated social support, more extensive media use, and a higher BMI" (Weaver 1). Throughout the studies the whole purpose was to figure out if there was a correlation between the amount of time a person played to the increase on their body weight. This hypothesis was introduced because experimenters believed that video games tend to lead to violent behaviors, obesity and muscular problems. | The whole purpose of conducting the experiments was to figure out if there was a correlation between the amounts of time a person played to the increase on their body weight. This hypothesis was introduced because scientists believed that video games tend to lead to violent behaviors, obesity and muscular problems. |

| | | | |
|---|---|---|---|
| 0 | I would try and revise the last sentence in the last paragraph so that it makes more sense and you can get your point across. | Playing video games causes bad health because there is not any physical activity evolved. Gaming only consists of sitting and pressing buttons. It can cause weight gain and it probably affects the eyes. From the lack of movement and the junk food that is most likely involved, adults will gain weight easier then juveniles due to the fact that their metabolism has slowed down since their teenage years. The fact that video gaming involves continuously looking at a screen, it most likely causes strain to the eyes. | Playing video games causes bad health because there is not any physical activity involved. Although there are a few types of interactive gaming devices, typical gaming only consists of sitting and pressing buttons. It can cause weight gain and it probably affects the eyes. From the lack of movement and the junk food that is most likely involved, adults will gain weight easier then juveniles due to the fact that their metabolism has slowed down since their teenage years. In turn, this study does not actually have any one correct answer to it because of the fact that casual conclusions cannot be formed from correlative studies. The fact that video gaming involves continuously looking at a screen, it most likely causes strain to the eyes. All I all, the overall outcome of gaming has more on a negative effect on the human body than a positive effect. |

| | | | |
|---|---|---|---|
| 1 | in the first sentence of the third to last paragraph: 'correlational studies cannot end in causal conclusions because the relationship between the two variables is always presently unknown.' what does 'the relationship between the two variables is always presently unknown' mean? | Correlational studies cannot end in causal conclusions because the relationship between the two variables is always presently unknown. The variables being studied may have the group studied in common, but that cannot prove whether one causes the other or even whether they relate at all. Correlated variables usually occur alongside each other in time, not one after another, thus there is no scientifically valid cause and effect. | Correlational studies cannot end in causal conclusions because the relationship between the two variables is always presently unknown. At any one period in time, variables with a correlational relationship can exist simultaneously. For instance, a gamer can be both depressed and fat before, during, or even after playing video games. In a correlational study, the variables are not dependent upon each other's existence in time, much unlike variables with casual natures. The variables being studied may have the 'group studied' in common, but that cannot prove whether one causes the other or even whether they relate at all. Correlated variables are capable of occurring alongside each other in time, and are not limited to one after another, thus there is no scientifically valid cause and effect on that basis alone. |