A validation study of students' end comments: Comparing comments by students, a writing instructor, and a content instructor

Melissa M. Patchan[1], Davida Charney[2] & Christian D. Schunn[1]

[1]University of Pittsburgh | USA

[2]University of Texas | Austin

Abstract: In order to include more writing assignments in large classrooms, some instructors have been utilizing peer review. However, many instructors are hesitant to use peer review because they are uncertain of whether students are capable of providing reliable and valid ratings and comments. Previous research has shown that students are in fact capable of rating their peers papers reliably and with the same accuracy as instructors. On the other hand, relatively little research has focused on the quality of students' comments. This study is a first in-depth analysis of students' comments in comparison with a writing instructor's and a content instructor's comments. Over 1400 comment segments, which were provided by undergraduates, a writing instructor, and a content instructor, were coded for the presence of 29 different feedback features. Overall, our results support the use of peer review: students' comments seem to be fairly similar to instructors' comments. Based on the main differences between students and the two types of instructors, we draw implications for training students and instructors on providing feedback. Specifically, students should be trained to focus on content issues, while content instructors should be encouraged to provide more solutions and explanations.

Keywords: content instructor feedback, writing instructor feedback, peer feedback

journal of WRITING RESEARCH

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. http://dx.doi.org/10.17239/jowr-2009.01.02.2

College students' writing ability does not seem to meet the standards expected by educators. For these students, the problem has compounded from elementary school through high school and into college. Since 1998, the writing aptitude of elementary and secondary students in the United States has been assessed using a common assessment allowing for longitudinal analyses (Salahu-Din, Persky, & Miller, 2008). Three achievement levels (*Basic*, *Proficient*, and *Advanced*) were proposed with the *Proficient* level as the baseline level that all students at a given age should be capable of performing. While some gains have been made in the past 10 years, the percentage of students in fourth-grade, eighth-grade, and twelfth-grade achieving a proficient level or better has been relatively low (only 28% for fourth-graders in 2002, 33% for eighth-graders in 2007, and 24% for twelfth-graders in 2007). Thus, many students are beginning college with only basic writing skills. Based on these statistics, one might wonder how writing instruction at the university level should be structured to help these students bring their writing ability up to standards.

The largest gains in writing ability seem to come from instruction that provides clear objectives through the use of specified criteria or scales, illustrates principles by working through concrete examples, and encourages students to interact in small groups (Hillocks, 1984). Such practices build upon learning theories that promote active learning, including collaborative and cooperative learning, provision of feedback, repeated opportunities to practice, and relevant domain-specific tasks (Ashbaugh, Johnstone, & Warfield, 2002; Palincsar & Brown, 1984; Prince, 2004). Therefore, the ideal writing assignment would involve realistic writing tasks, multiple drafts to allow for planning and revision, and feedback from readers. However, such rich assignments are not easy to implement systematically in college classes that are not dedicated exclusively to writing instruction. Grading writing assignments requires considerable effort, especially when feedback on how to improve writing is to be provided. Typically, neither instructors nor teaching assistants have any training in teaching writing or providing feedback.

As class size increases, the instructor's ability to incorporate writing assignments diminishes. In order to include writing assignments in courses that would otherwise have them, many instructors have considered peer-review as a supplement or alternative to instructor feedback. However, many instructors are hesitant to use it in their classroom. One of instructors' biggest concerns has been whether students are actually capable of grading their peers' papers accurately and responsibly. Several studies found that, under appropriate circumstances, students are able to provide reliable and valid ratings of writing (Cho, Schunn, & Wilson, 2006). These studies support the use of student ratings for grades, which then allows for more opportunities to write, even in large classes.

However, the question remains whether the peer-review process can actually improve student writing. Are students capable of providing feedback that is at least equivalent to the feedback provided by an instructor? This question is especially important in smaller classrooms that already include writing, where the purpose of the

writing assignments is to help students learn discipline-appropriate writing conventions. Here the question of assessing paper quality is not as critical as the usefulness of the feedback. The current paper provides a first step in determining whether peers are able to provide helpful feedback that could improve students' writing ability.

In order to understand the nature of student feedback, a good first step is to compare student feedback and instructor feedback. It is also important to consider more than one source of instructor feedback available to students: feedback from writing instructors and feedback from content instructors.

1.1 Feedback from Writing Experts

A primary source of instructor feedback is the first-year writing course, which is generally required of students with low scores on standard aptitude tests. These students typically amount to about half the incoming class at a major public university. First-year courses at two- or four-year colleges are usually small (less than 25 students) and focus on developing an understanding of the various uses of writing (such as persuasive, descriptive, narrative, critical, etc). These courses are designed to incorporate the features viewed as important in improving writing ability with multiple drafts and feedback from the instructor. Typically the instructor of a first-year writing course has or is pursuing an advanced degree in English and has access to a wealth of scholarly and pedagogical resources for teaching writing processes. When asked to comment on texts outside their domain of expertise (such as assignments in technical or business writing), writing instructors tend to focus on how to solve problems with the elements of the prose (such as coherence, organization, or appropriateness to audience) rather than the accuracy and comprehensiveness of the content (Smith, 2003a, 2003b).

Writing instructors are known to adopt a variety of roles when reviewing a student's writing, including a judge, coach, or a typical reader (Dragga, 1991; Fife & O'Neill, 2001). A judge's comments are likely to focus on problems, whereas a coach's comments are likely to focus on solutions. An instructor adopting either of these roles is more likely to provide explanations than someone who has adopted the role of a typical reader. While it is possible for composition instructors to adopt various roles, Smith (1997) found that the majority of their feedback was evaluative (72%), rather than coaching (20%) or a reader response (8%).

1.2 Feedback from Content Experts

In an effort to provide more writing instruction for students, many universities have established Writing Across the Curriculum (WAC) programs and Writing in the Disciplines (WID) programs. WAC programs use writing to promote the learning of disciplinary skills and content, while WID programs focus on the learning of discipline-specific discourse practices (Ochsner & Fowler, 2004). As a result of these programs, writing instruction occurs in courses where the instructors are first and foremost experts in the content rather than in writing instruction. As writers, the content instructors may have a lot of experience from writing their own papers, but they are probably less

aware of the elements of writing and writing processes because they have automatized the writing process. Therefore, content instructors are more likely to focus their feedback on problems with the content and the disciplinary genre (Smith, 2003a, 2003b).

Even with the introduction of WAC and WID programs, the number of writing opportunities available to students is very limited. For example, we found in a recent review of syllabi from psychology departments at 12 national universities of varying rank, some psychology departments at both large and small institutions never require paper drafts. This problem brings us back to the need for an increase in the use of peer-review, and using peer-review leads to the source of feedback that the current study is most concerned about: the student.

1.3 Feedback from Student-Peers

Student feedback may resemble either that of a writing instructor or that of a content instructor. Relative to a writing instructor, students may know more about the content discipline and the particular content questions being examined. On the other hand, relative to the content instructor, they have less practice at discipline-specific writing and less familiarity with disciplinary genres. Therefore, if asked to provide feedback to others, they may mimic the feedback on writing assignments in high school English or in their first-year writing course.

1.4 Feedback Features

Only one prior study has explicitly compared student comments to instructor comments. Cho, Schunn, and Charney (2006) found that a content instructor provided more feedback than did students when asked to comment on the same three writing dimensions on a sample of the same student papers. Both in a large survey class and in a small disciplinary class, feedback provided by a content instructor contained more words and raised twice as many ideas as the feedback from students.

Cho et al. (2006) also found that students included more praise than the content instructor. It is possible that the content instructor appreciated students' papers less, had less understanding of the value of praise for motivating revision, or did not think praise was a core part of the genre of comment giving. However, more research is needed to distinguish which explanation is most accurate.

Finally, Cho et al. (2006) found that the content instructor used more directive comments than did students. In directive comments, a reviewer suggests context-specific changes that may not apply to any other papers. The directive comments appear to be important for instigating revision activity, but Cho et al. did not differentiate sufficiently among different qualities of directive comments.

The study reported here was designed to differentiate subtypes of directive comments and compare their frequency in the feedback of instructors and students. We focused on dimensions of feedback that have been thought to be important in improving writing. These included whether summaries, identified problems, suggested

solutions, locations, explanations to problems, explanations to solutions, global issues, local issues, praise and/or mitigating language (Nelson & Schunn, 2009).

We focused on two features of directive comments: the explicitness with which a problem or a solution was described and the degree of explanation that was offered. A problem/solution description may be as simple as a label for a problem (e.g., "confusing paragraph" or "wrong format") or a solution ("put this sentence first" or "use APA style"). We categorized comments identifying problems and solutions separately because, overall, detection of problems has been found to be easier than articulating solutions (Flower, 1986). Further, the detection ability of instructors and students may differ as well as their ability to describe the problem clearly. As a result of their greater experience and relevant knowledge, content instructors are most likely to notice problems related to the topic and details of the content. However, writing instructors may be more experienced at detecting writing problems and more successful at describing a problem clearly (Smith, 2003a, 2003b). As for solutions, because of their greater practice at writing comments and their knowledge of effective revision techniques, writing instructors may provide more explicit solutions than either content instructors or students.

The second feature we examined was the presence of explanations of problems and/or solutions. Explanations concern why a problem decreases the quality of the paper or why a solution makes the paper quality better. Explanations regarding a problem may arise for a number of reasons from having information that the writer may lack to feeling uncertain about the problem. Explanations of a solution, however, relate to the role a reviewer adopts *vis à vis* the writer. Some reviewers see their duty as correcting the text rather than guiding the writer; such reviewers simply issue imperatives or even edit a student's paper directly (Sperling & Freedman, 1987). A reviewer who explains a solution is allowing the writer to take responsibility for making decisions about the text. Some instructors offer explanations to more capable students but edit the papers of weaker students (Herrington, 1992). Students may offer explanations to a peer either out of respect for their equal status or a sense of uncertainty about whether the solution is a good one.

One thing to consider when examining feedback differences between peers and instructors is whether guidelines were provided regarding how the feedback should look. For example, consider students' ability to focus on global problems during revision. Typically, novice writers have a lot of difficulty with this task. However, students who were instructed to focus on global issues during revision made more global changes than students who were instructed to improve their text (Wallace & Hayes, 1991). What makes this study especially impressive was that the students only received eight minutes of instruction on making global revisions.

In the current study, the peer-review process is scaffolded through the use of the "Scaffolded Writing and Rewriting in the Discipline" (SWoRD) system, which provides rubrics for reviewing and incentives to take the peer-review task seriously (Cho & Schunn, 2007). Therefore, it is possible that the scaffolded reviewing could influence

the types of comments produced. However, because all participants (i.e. the students and the instructors) received the same guidelines, there should not be systematic differences between groups that come as a result of differential guidelines. On the other hand, the use of such structures might reduce differences between groups relative to more unstructured reviewing.

1.5 Hypotheses

As a first step in determining whether the comments students provide are valid, we examined how various types of reviewers' (i.e., content instructor, writing instructor, and students) comments differ. Expectations regarding feedback features were as follows:

1. Content instructors are expected to identify more problems than writing instructors.
2. Writing instructors are expected to offer more solutions than content instructors.
3. Content instructors are expected to provide more problems regarding content, while writing instructors are expected to provide more problems regarding writing.
4. Instructors are expected to produce longer comments than students.
5. Students are expected to include more praise than instructors.
6. In general, students are likely to sit between content and writing instructors.

2. Method

2.1 Overview

Comparing students' comments to writing and content instructors' comments is difficult because students typically receive comments from only one of these sources on a given text, and especially not from all three at once. If we had simply sampled naturally occurring feedback from these sources, we would not know whether the differences were due to the feedback giver, the writing assignment, or the writing approach taken by the author given different evaluative audiences. Our approach, instead, was to collect both first and second drafts of undergraduate writing, as well as the comments produced by peers regarding how to improve the first draft. After the course was completed, a content instructor and a writing instructor were paid to produce a review with comments for each of the first drafts the students reviewed. We then compared the proportion of various feedback features included in comments by three types of reviewers: content instructor, writing instructor, and students.

2.2 Course Context

The class chosen for this study was a large undergraduate survey course entitled, *History of the United States, 1865-present*. The course satisfies the university's history requirement. Students were required to write a six-to-eight page argument-driven essay responding to one of two possible topics: (1) whether the United States became more democratic, stayed the same, or became less democratic between 1865 and 1924, or

(2) examine the meaning of the statement "wars always produce unforeseen consequences" in terms of the Spanish-American-Cuban-Filipino War and/or World War I. A large portion (40%) of the students' course grade was based on the performance of the writing assignment. The instructor did not comment or grade the papers during the course, but instead students used the SWoRD system to receive comments and grades from their peers.

2.3 Participants

The student participants were enrolled in this large undergraduate history course. The majority of the participants were male (62%), Caucasian (73%), and between the ages 18 and 21.

The content instructor (CI) was the instructor of this course, an associate professor of history who has taught this course many times. She was also very committed to writing in the disciplines, having coordinated the department's substantial writing seminar and having participated in faculty development seminars on Writing Across the Curriculum.

The writing instructor (WI) was a skilled instructor of college-level writing with graduate training in rhetoric and composition. She was employed as a full-time visiting lecturer in the university's English department. She also was the Outreach Coordinator at the university's writing center, and she has served as the director of the Freshman Engineering Integrated Curriculum, which focuses on incorporating writing within the engineering major.

When comparing student and instructor comments, it is important to note that expertise is a continuum, rather than the dichotomy of one having expertise or not. We purposely chose cases that are clear examples of expertise, although not the most extreme cases of expertise in either writing or content. Therefore, other instructors with more or less expertise will likely have the observed characteristics to a greater or lesser extent than what we observe. All of the students may be considered novice writers because the writing genre would likely be new to them. However, students' writing expertise is also likely to range across a continuum. For the current study, we addressed possible variations in student writing skill level by also dividing students into two skill levels based on the grades they received on their first draft of the paper assigned in this class: low peers (i.e., students with weaker writing skills) and high peers (i.e., students with stronger writing skills).

2.4 Review Support Structures: Review rubric and SWoRD system

The peer-review process was facilitated using SWoRD, an anonymous web-based reciprocal peer-review system. Students submitted their first draft online, and the system automatically distributed them to six other students. The students then had two weeks to review their peers' papers. They were required to provide comments and a rating on three dimensions: prose transparency, logic of the argument, and insight beyond core readings. *Prose* transparency focused on whether the main ideas and transitions between ideas were clear. *Logic* of the argument focused on whether the paper

contained support for main ideas and counter-arguments. *Insight* focused on whether the paper contained new knowledge or perspectives beyond the given texts and course materials. Appendix A shows the exact reviewing rubric used.

The two instructors were provided with hard copies of the papers and were asked to review them as they would for one of their own classes. They were also provided with the same reviewing prompts as the students.

The SWoRD process described above is very similar to many journal, conference, or classroom peer-review systems. It is important to note that SWoRD has two additional features beyond the typical system, which focus on student accountability: automatic rating accuracy measures and author-generated review helpfulness ratings. These features were designed to improve the quality of peer reviews.

Half of a student's reviewing grade is based on the accuracy of their ratings. SWoRD automatically evaluates the accuracy by calculating the reliability of each student's ratings in terms of systematic differences (i.e., the extent to which the student assigns either all high ratings or all low ratings), consistency (i.e., the extent to which the student is able to distinguish good papers and bad papers), and spread (i.e., the extent to which the student appropriates uses the full range of possible ratings). Then these reliability scores are summed in order to generate the student's rating accuracy score. Basing a portion of the student's reviewing grade on their accuracy is likely to lead to more attention to the reactions of other students, and thus to a more general audience than just themselves.

The other half of a student's reviewing grade is based on the author-generated helpfulness rating. After submitting the revised draft of the paper, the students rate how helpful they found each reviewer's comments on a scale of 1 to 7. This rating from each of the papers that the student reviewed is used to determine the helpfulness score. By using author-generated helpfulness ratings, students are more likely to form revision-oriented comments. Because students only rated peer feedback in this way, these ratings do not serve as data in the current study.

2.5 Selected Papers

To make the coding process manageable while maintaining sufficient statistical power, we selected 24 papers from the 111 registered students for analysis. The majority of the papers were randomly selected, but to make sure that we had some papers that received relatively useful feedback and relatively week feedback, we selectively picked eight papers. For these eight papers, we selected papers that started out around the same quality (i.e., similar first draft scores), but had either relatively high gain scores (4 papers) or relatively low gain scores (4 papers). By selecting papers with similar scores, the changes were more likely attributable to the feedback than the author. Each of the 24 papers received comments from on average six peer reviewers and each of the experts, generating 187 reviews. Because peer reviewers are randomly assigned, these reviews came from 74 different students in the class.

2.6 Coding Process

The data analysis examined the relative rates of comment types at several different levels of granularity. To help make clear how the levels relate to one another, Figure 1 presents a schematic of the data organization. Each review contained comments in each of the three dimensions, but a single reviewer's comments on a given dimension often contain several different ideas or suggestions. Thus, the comments within dimension were separated into separate coherent segments. Because students often had comments under one dimension that related to another dimension, we combined all the segments together within a review, ignoring dimension differences in our analyses. Thus, we have some number of total segments for each review, varying from 4 to 11, with a mean of 7.2. Some of our analyses were at this total review level (e.g., how many total segments per review or how many words per review).

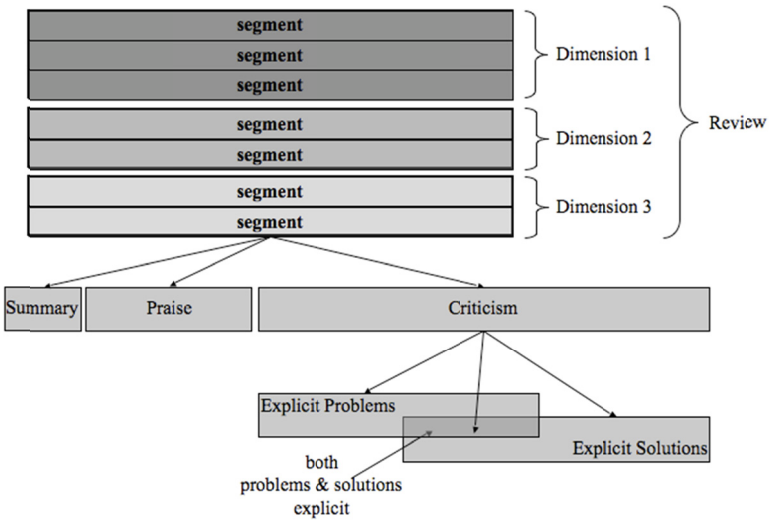

Figure 1. The relationship between the levels of analysis of the comment data.

After segmentation, each of the comments was double-coded into various categories. Two judges coded the segments independently. When there was disagreement between the judges, they would discuss their choices and decide on the most appropriate code. The inter-rater reliability ranged from moderate to high Kappa values of .51-.92 (Landis & Koch, 1977).

First, segments were divided into the categories of summary comments, praise comments, and criticism comments (see Appendix B for definitions, examples, and kappas). We counted how many segments of each type occurred. Then we coded features of each of those types of comments. For example, of the criticism comments,

how many were easily localized within the document, or how many included mitigation language?

Then, critical comments were divided into ones that explicitly mentioned what the problem was versus ones that left the problem implicit (i.e., provided only a solution). Critical comments were also divided into ones that explicitly mentioned solutions versus ones that left finding a solution up to the author. These two dimensions were independent, where some critical comments could have both a problem and a solution (as indicated in Figure 1).

Finally, our coding delved into the nature of the feedback. These codes can be categorized into two areas of focus: affective language and specificity. In addition to praise, the criticism comments were also coded for the type of affective language used. This area of focus distinguished whether mitigation, inflammatory language, or neutral language was used. Within the mitigation classification, whether the mitigating language included compliments, questions, or downplay was also indicated.

The second focus was on specificity, building on the general differences in directive language found by Cho et al. (2006). The specificity of both the praise and the criticism comments was coded. This specificity included whether the location of the problem/solution was indicated and at what level the scope of the problem/solution was (i.e., word/sentence level, paragraph/part-of-paper level, or whole paper). In addition, certain features of specificity were also coded for problems and solutions separately: whether explanations were provided and the focus of the feedback (i.e., whether the problem and/or solution focused on the high prose, low prose, or the substance of the paper). Summaries were also coded as either those that focused on claims that the author made or some other type of summary (such as stating what the writer did without any criticism or praise).

3. Results & Discussion

3.1 Overview

We present the results in four layers: 1) differences between the content instructor and the writing instructor; 2) differences between the peers and both kinds of instructors; 3) situations in which the peers were similar to one of the two instructors or fell between the two instructors; and 4) an analysis of how well these patterns held across peer commenters of high and low writing ability and across papers of high and low quality.

To examine the statistical reliability of observed differences, we treated documents as the object of variability. That is, we calculated a mean rating for each document for each rater type (e.g., for paper 1, the mean number of comments per student review, the number of comments from the content instructor, and the number of comments from the writing instructor). Then we examined the consistency in differences of values across groups taking into account the variability across papers. It was necessary to use

paper as the object of variability rather than rater because we only had one content instructor and one writing instructor.

We conducted statistical tests (within-subject ANOVAs, with Bonferroni-Dunn post hoc comparisons to establish pair-wise differences between groups) on all of the measures that were coded. Appendix presents all the means for each group on each dimension for full disclosure of the observed results. To avoid filling the text with inferential statistics and null results, we use the convention that all results reported as significant effects were at least $p < .05$ and all results treated as non-significant were $p > .05$. On each graph that can be used to infer approximate statistical differences, we include error bars based on RMSE because it is more appropriate for within-subject analyses than standard error bars (Loftus & Masson, 1994): if the error bars overlap, the difference is not likely to be statistically significant.

To provide a better understanding of the semantic significance of the effects, we use Cohen's *d*, which is the difference in the means between two cases divided by the average standard deviation within each case (Cohen, 1988)—in other words, how many standard deviations are the two means apart from each other. We consider an effect to be small when *d* is .2, medium when *d* = .5, and large when *d* = .8 or greater. Note however, that theoretically *d* is not capped at 1; one can find differences between means that are two or more standard deviations apart. Indeed, in our context, we will see that some of the differences between groups are very much larger than just *d* = 1.

3.2 Differences Between the Two Instructors

Overall, the instructors generated a similar quantity of feedback to student papers. The writing instructor (WI) and content instructor (CI) did not significantly differ in the number of feedback segments (means CI = 7.4, WI = 6.4) or the number of words in the feedback (means CI = 369, WI = 347). The similarity allows us to analyze the more specific comment types without distinguishing between comment absolute frequency and relative frequency. Moreover, the two instructors generated similar numbers of praise statements, summary statements, and critical statements. Where the instructors began to differ from one another was within the critical statements.

Critical comments may be explicit in identifying a problem, a solution, or both. That is, a comment may explicitly describe the problem (e.g., "the transition between the third and fourth paragraphs was rough") or leave it implicit by only giving a solution. Similarly, critical comments may explicitly provide a solution (e.g., "add a transition sentence") or not, leaving the writer to find a solution.

As expected, the two instructors differed dramatically in terms of what they described explicitly (see Figure 2A). The content instructor typically was very explicit about the problem, whereas the writing instructor only described the problem explicitly about 25% of the time. By contrast, the writing instructor almost always gave an explicit solution, whereas the content instructor only gave an explicit solution 50% of the time. The instructors paired an explicit problem with an explicit solution infrequently but equally often (CI = 35%, WI = 23%).

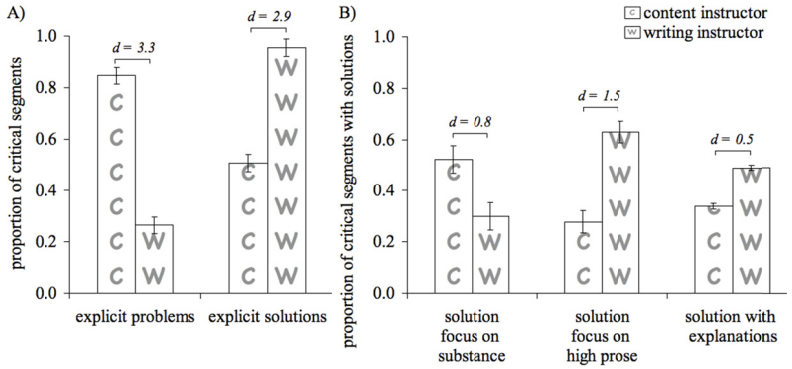

Figure 2. A) The extent to which the content and writing instructor gave explicit problems or solutions in their critical comments. B) The extent to which the content and writing instructor focused their explicit solutions on issues of argument substance in history versus on issues of high level prose problems.

In addition to differing in how often they provided explicit solutions, instructors also differed dramatically in terms of the content of the solutions (see Figure 2B). The content instructor's solutions discussed substantive issues (e.g., historical facts or relevant historical trends) over 50% of the time, whereas the writing instructor's solutions did so only 30% of the time. By contrast, the content instructor's solutions only involved high-level prose issues (e.g., argument flow or problems with transition) 30% of the time, whereas the writing instructor's solutions did so over 60% of the time. The instructors rarely gave low-level prose solutions or solutions that focused on assignment-specific issues.

Another way to think about this last factor is not in conditional terms (i.e., given a solution, what was the content?) but rather in absolute terms (i.e., how often did feedback contain a given solution type?). Because the writing instructor was much more likely to give an explicit solution, the absolute frequency results are different from the conditional results. In absolute terms, the content and writing instructor provided a similar number of substance solutions (CI = 0.8, WI = 1.2), but the difference in high-level prose solutions was much greater than the conditional analysis would suggest (CI = 1.2, WI = 2.4).

The content of solutions relates to the presence or absence of explanations of those solutions. The instructors differed on this dimension as well. Consistent with their overall solution focus, the writing instructor gave explanations for explicit solutions almost 50% of the time, whereas the content instructor only did so a third of the time.

It is worth noting that while the content of the solutions did differ, the content of the explicit problems did not differ. That is, when instructors explicitly described problems,

they were equally likely to comment on substance problems or high-level prose problems, and they were equally likely to explain problems. The total number of such comments differed, however, reflecting the large differences in whether problems were mentioned explicitly or not.

The last significant difference between the two instructors involved mitigation in critical comments. We coded for three different ways of providing mitigation: including a compliment with the criticism, turning the criticism into a question, and downplaying the importance of the criticism. The writing instructor was slightly more likely to include any form of mitigation (CI = 23%, WI = 34%), but this overall difference was not statistically significant. The real difference concerned compliments: the writing instructor complimented students significantly more often than did the content instructor (CI = 6%, WI = 17%, d = 0.7).

For the most part, these results matched our expectations. Previous research has shown that content instructors were more likely to comment on problems (Smith, 2003a, 2003b); writing instructors' training and expertise raises the likelihood that they will offer solutions and explain them. Interestingly, while we expected to find differences in the content of problem explanations, these did not emerge.

3.3 Consistent Differences Between the Peers and Both Instructor Types

Next we turn to differences between peers and the two instructors (i.e., the peers differed in the same way relative to both instructors, either higher than both instructors or lower than both instructors). Overall, the peers (P) and the instructors (I) did not differ in terms of the overall number of comments per review provided (means P = 7.8, I = 6.9). The instructors had a significantly higher workload commenting on 24 papers instead of just six, but they also were given more time to complete their task. In other circumstances, the instructors might have produced more or fewer comments. Analytically, however, this lack of differences allows us to ignore differences at this grain-size between the absolute number of subtypes and the relative frequency of subtypes of comments.

Although the total number of comments per review did not differ, the number of words per review did: peers generated fewer words per review than did instructors (means P = 272, I = 358, d = 1.5). In other words, the instructors did not have more comments; they just used more words to make those comments.

Comments were divided into summary, praise, and criticisms at the next level of analysis. Here a consistent difference was that peers generated almost twice as much praise as the instructors (see Figure 3A), similar to the findings of Cho et al. (2006).

Logically, one might expect that a difference in amount of praise combined with no differences in total number of comments would imply differences in amount of summary or critical comments. However, these differences were not consistent across instructor type. Because the instructors themselves differed so much, the peers did not differ from the two instructors in a consistent way in terms of providing explicit problems or explicit solutions.

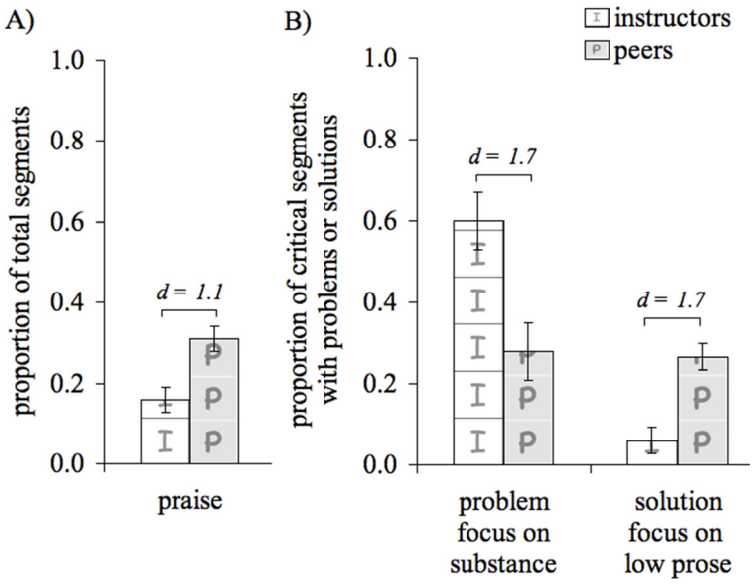

Figure 3. A) Differences between instructors and peers comments overall on tendency to include praise comments. B) Differences between instructors and peers on whether explicit problems tended to involve substance issues and whether solutions tended to address low prose issues.

Consistent differences between instructors and peers did emerge in the content of the problems and solutions (see Figure 3B). When they explicitly mentioned problems, the instructors were twice as likely as peers to focus on substance. It is particularly interesting that the writing instructor focused more on substance problems than the peers because presumably the peers had more access to historical details relevant to their papers. Perhaps the peers felt less authorized to comment on issues of historical content or the writing instructor was more skilled at using content within the paper to determine if there were inconsistencies (Smith, 2003a, 2003b).

With respect to solutions, peers were more likely to provide low-prose solutions than were instructors. Note that no one focused on low-prose issues often, reflecting the emphasis placed in the commenting rubric NOT to discuss low-level prose problems; peers were simply more likely to ignore the rubric instructions and provide some low-level solutions.

Another consistent point of difference between instructors and peers involved mitigation language in criticisms. The difference is not in the overall amount of mitigation, but rather the difference related to downplay mitigation in particular.

Instructors never used downplay mitigation (i.e., they never made a criticism and then noted that the problem might not be so important), whereas peers did occasionally use this form of mitigation. However, even the peers rarely downplayed (mean P = 2.5%, *d* = 0.9).

Again, these results are consistent with our predictions. Finding that instructors were more verbose than the students replicates the findings of Cho et al. (2006), as did finding that students provided more praise than instructors. However, unlike Cho et al. (2006), we found no differences in the number of ideas in the reviews. These differences between contexts are not unexpected, because the instructors' workload would be different. There were some other instructor-peer differences, but those differences were not consistent across instructor type and will be discussed in the next section.

3.4 When Peers Fall between Instructor Types

For each of the dimensions for which there were differences between the two instructors, the question arises: are the peers more like one of the two instructor types or somewhere in between? As indicated by Figure 4, the answer to that question depends upon the feedback feature.

The peers sit between the two instructors when it comes to providing explicit problems and explicit solutions; unlike the instructors, the peers were equally likely to mention problems and solutions (see Figure 4A). This pattern also holds for whether or not solutions focus on high prose issues (see Figure 4C).

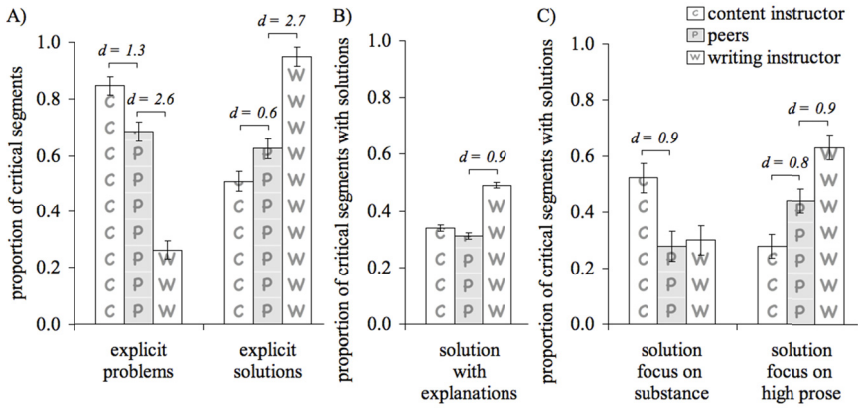

Figure 4. Location of peers relative to differing instructor patterns on A) presence of explicit problems or solutions, B) inclusion of explanations for solutions, and C) focus on substance versus high prose issues in solutions.

Peers resembled the content instructor in omitting explanations to solutions (see Figure 4B), and on the other hand, they resembled the writing instructor in omitting solutions

on substance (see Figure 4C). In both cases, the peers do less than the maximum possible amount of the activity but at least do no worse than one of the two instructors.

On only one dimension did peers resemble the more active instructor: the use of compliments to mitigate the criticisms. Peers and writing instructors gave higher proportions of compliments than the content instructor (CI = 6%, WI = 17%, P=19%).

3.5 Dimensions with No Differences

Despite the above-mentioned differences, peers and instructors were relatively similar in the comments they provided on student papers. The degree of similarity can be assessed by calculating the correlation between each pair of commenter-types in the relative use of each of the feedback features. To make this calculation, we developed a commenting profile for each commenter type that included the proportion of use for each feedback feature. In order to eliminate dependency, one of each set of codes was removed from the analyses. Then a (Pearson) correlation of these commenting profiles was calculated between each pair of commenter-types (i.e., peers vs. writing instructor; peers vs. content instructor; writing instructor vs. content instructor). We found that the commenting profiles were generally very similar. In fact, the peer profile correlated with each of the instructor profiles (r = .80 with CI and r = .69 with WI) to a greater extent than the two instructor profiles correlated with each other (r = .66), although all the correlations were of a similar magnitude.

The general similarities in the profiles were in part a consequence of the number of feedback features with no group differences (i.e., feedback features for which peers and both types of instructors performed in very similar ways). It is worth summarizing what these feedback features were to highlight the range of dimensions for which peers' feedback seems similar to instructor's feedback. The dimensions were: 1) the number of critical comments, 2) the number of summary statements, 3) the frequency of providing both explicit problems and solutions for a critical comment, 4) the tendency to explain problems, 5) the overall use of mitigation, 6) the likelihood of clearly giving the location of a problem in the document, 7) the level of the problem being discussed (word vs. paragraph vs. greater level), and 8) the likelihood of pointing out low-level and high-level writing problems explicitly. One may wonder about the effect sizes of some of these non-significant relationships. While there were some moderate effect sizes (e.g., proportion of critical comments, proportion of questions, proportion of high prose problems), the majority of the effect sizes (75%) were 0.4 or smaller. Some of the instances of non-significant, moderate effect sizes could be a result of power issues. More specifically, sometimes when problem and solution comments were further categorized (e.g., high prose, low prose, substance), there would be no data for one of the instructors because they either provided only problems or only solutions to a particular paper. For these categories, if we had more data points, there may have been other significant effects. Despite these power issues, peer feedback seems to be quantitatively and qualitatively similar to instructor feedback, at least in this setting.

3.6 Interactions with Paper Quality and Reviewer Writing Skill

Because these data came from a particular class setting, it is unclear whether the patterns would hold up across settings in which the peers had lower writing skills (e.g., in a more introductory course or a remedial writing skills course) or higher writing skills (e.g., in a more advanced course or in an university with a stronger writing training program). Similarly, it is unclear whether the patterns of group differences observed thus far are specific to papers of a particular quality level or whether these differences would hold across papers of higher and lower quality.

Due to the broad cross-section of students who take survey courses, we had sufficient range in paper quality and reviewer writing skill within this course to allow us to pursue these questions to some extent. We examined whether the patterns hold for both higher and lower quality papers (using a median split on average rating of final drafts) and for peers with more and less writing skill (using a median split on their mean paper grades as writers). These interactions were assessed formally through two-way ANOVAs, adding paper quality (between subjects) or reviewer writing skill (within subjects) to the existing one-way ANOVAs.

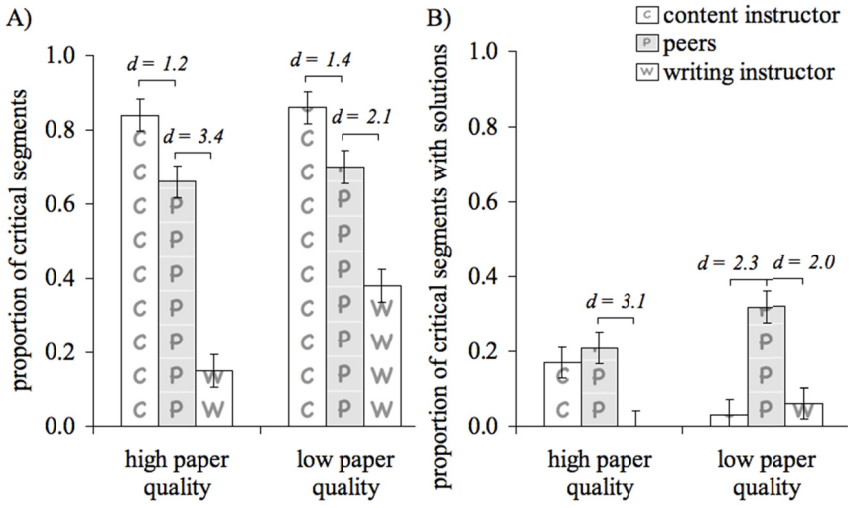

Figure 5. Separation of high and low quality papers for A) the proportion of critical statements with explicit problems and B) the proportion of explicit solution statements that focus on low prose issues.

Regarding paper quality, there were only two significant interactions. The first involved the likelihood of providing an explicit problem in a critical statement (see Figure 5A). This interaction was a simple quantitative interaction. The pattern for the content instructor and the peers was the same for high and low quality papers; while the writing

instructor was much less likely to provide explicit problems for both quality levels. The likelihood was especially low for high quality papers.

The second interaction involved the focus on low prose issues in solutions (see Figure 5B). Peers were the most likely to provide any low prose solutions, and they did so for both high and low quality papers. While the content instructor provided about the same proportion of low prose solutions as the peers for the high quality papers, no low prose was offered for the low quality papers. The writing instructor omitted this type of comments altogether.

In examining interactions with reviewer writing skill, only one significant interaction was found. Recall that peers were the only ones who downplayed a problem. One possible interpretation of that result was that peers were unsure of the problems they raised. However, it was only the higher skilled peers that used this type of mitigation (mean = 5%). Thus, this strategy was more likely to be a matter of rhetorical style in commenting by peers rather than one of being unsure of the comments being made.

Overall, we see very little change as a function of peer writing skill or paper quality. Thus we have some confidence that these results would generalize to other courses or university setting with students at different writing levels.

4. General Discussion

4.1 Summary of Findings

The current study provided a detailed examination of how valid student comments were; that is how similar students' comments were to instructors' comments. As one might expect, instructors' comments were longer than the students' comments. However, they both commented on the same number of ideas. More interesting was how instructors and students differed in the feedback features.

First, there were differences in the type of affective language used. Students used praise twice as often as the instructors, and both the students and the writing instructor used compliments with their criticisms almost three times as often as the content instructor. Instructors never downplayed their criticism, but students did use this strategy on rare occasions. Furthermore, the high-skilled peers actually used downplay more than the low-skilled peers, suggesting that the use of downplay does not reflect uncertainty in the advice but rather mitigation of negative emotional impact of criticism.

Second, the instructors and students differed in the rate and content of explicitly identified problems. As expected, the content instructor was especially problem focused. Overall, the content instructor explicitly identified more problems than the students, who did so more than the writing instructor. The writing instructor explicitly identified even fewer problems on high quality papers, suggesting perhaps that the writing instructor thought problems needed only to be explicitly identified if they were

not likely to be obvious to the author. In addition to rate differences, both instructors focused the problems identified on issues regarding the content of the paper more so than did the students.

Finally, there were rate and content differences in the explicit solutions that were offered by the instructors and students. Overall, as expected, the writing instructor was most solution-oriented, offering more solutions than the students, who did so more than the content instructor. Not surprisingly, the content instructor provided the most solutions regarding the content of the paper, the writing instructor provided the most solutions regarding high prose, and students provided the most solutions regarding low prose. Interestingly, students' focus on low prose solutions increased on lower quality papers, while the content instructor's focus on low prose solutions increased on the higher quality papers—they clearly had different theories about what kinds of problems to emphasize with weaker writers. In addition, the writing instructor provided the most explanations to the solutions.

4.2 Why Do Students Provide More Praise?

We were able to replicate Cho et al.'s (2006) finding that students provide twice as much praise as instructors. It is clearly a strong phenomenon; but it has not yet been explained. There are at least three possible explanations for this pattern: 1) students may be more impressed by their peers' papers than are instructors, 2) students may think that praise is an important part of the commenting genre, or 3) students may appreciate receiving praise so they provide it when they comment. There is some evidence that the third explanation is most likely to be the underlying cause.

If students provided more praise because they were more impressed by the papers, there would likely be general differences by student ability; that is low-skilled students would be providing more praise than the high-skilled students and higher quality papers would receive more praise than lower quality papers by both students and instructors. However, the amount of praise provided was not driven by the skill level of the reviewer (i.e., low-skill students did not provide more praise) nor was the amount of praise driven by the quality of the papers (i.e., higher quality papers did not receive more praise).

Another possibility is that students have a template of how they think feedback comments should look, and included in that template is praise; whereas instructors may have their own template for feedback comments in which praise plays a lesser part. To test for this kind of template explanation, we examined the order in which praise was provided. If praise were part of a feedback template, it likely would be consistently at the beginning or end of feedback. Praise was generally associated with being first or last overall, however to differing degrees across participant groups. In particular, the writing instructor was the most likely to place praise in a very particular location ($p < .0001$), and the content instructor and low skilled peers had a similar, but much more weakly expressed preference ($p < .1$ and $p < .07$, respectively). However, high skilled peers did

not have this tendency at all. Thus, the association of praise with student comment templates does not seem a strong explanation for the increase in use of praise.

However there is evidence that students strongly appreciate praise. For example, Cho et al. (2006) found that students were significantly more likely to rate the feedback that they received as helpful if some praise was included in the comments. This appreciation of praise by students may thus be the best explanation why they are especially likely to include praise when they provide feedback to their peers.

4.3 Implications for Practice

Implications for WAC/WID: Challenges for Content Learning. Instructors in disciplinary areas are more likely to provide comments that encourage more engagement with the subject matter, or they will at least correct the misconceptions that they see in student papers. This type of comment is not as likely to be provided by student peers. Therefore, instructors using peer review to promote writing-to-learn will need to develop additional support to focus peers on content issues.

Implications for WAC/WID: Improving Writing Skills. Disciplinary instructors are quite capable of detecting and explaining problems with students' writing, although their ability may well depend on their previous experience at assigning writing. On the other hand, their comments are much less likely than those of writing instructors to propose and explain solutions. Therefore, faculty development workshops should suggest to faculty that they spend a greater amount of attention to solutions, as well as perhaps more use of praise.

Implications for peer-review: Overall, students' comments are less verbose than those of instructors, but the number of individual ideas seen as worthy of comment is about the same. Under certain conditions, students provide comments that are similar in both quantity and quality to those of instructors. The relevant conditions in this study were likely the provision of a rubric and strong incentives to take both the writing and reviewing tasks seriously. This finding is especially striking since it applies to students whose own writing scores fell below the class median—being able to provide similar types of feedback as instructors does not appear limited to strong writers.

4.4 Caveats

A few threats to the generalizability should be addressed. First, this study's peer-review process was done online anonymously, and other methods of peer review (e.g., in face-to-face feedback) may change the distribution of peer feedback produced. For example, instructors may be more likely to include some praise when speaking directly to a student. Also in face-to-face settings, the author has an opportunity to ask for elaborations, so the instructor or peers may provide more explanations or solutions when prompted. In addition, feedback in the form of end comments may not generalize to marginalia. For example, high prose issues might be less commented upon in marginalia given that they involve a larger scope than just a sentence or paragraph. Also, praise and mitigation may be used less because most praise tends to be general

rather than about specific parts of the paper. Finally, there may be differences between the instructors' feedback that we collected (i.e., after the semester with ample time to spend on commenting) and an instructor's feedback that is provided with rapid turnaround and other pressures during the semester. The instructor's immediate feedback is likely to be shorter, and possibly focus even more on their discipline (i.e., a content instructor may focus more on the content and a writing instructor may focus more on the writing).

This study focused on the form of the comments rather than the quality of the comments. One reason for not focusing on quality was the difficulty in determining what constitutes high quality. One possible method would be to have instructors rate the comments generated by students and instructors on how useful they would be for a student. Cho et al. (2006) used such a method and found that instructors thought that another instructor's feedback was more helpful than students' feedback. However, there also has been evidence that instructors may not be able to accurately judge what students will find intuitive or difficult (Nathan & Koedinger, 2000). In fact, one study found that students did not rate instructor's feedback as more helpful than students' feedback (Cho et al., 2006). A recent dissertation examined the effect of peer versus student feedback on writing quality (Cho, 2005). Students were randomly assigned to one of three conditions: receive feedback from a single instructor, receive feedback from a single peer, or receive feedback from a group of six peers. The students were unaware of whether or not their feedback came from an instructor or peers. Interestingly, the papers written by students who used feedback from a group of peers improved more than those who revised using feedback from an instructor. Based on a path analysis, three possible explanations may account for the improvement differences: (1) peer comments included more praise, which appeared to increase the implementation rate; (2) peer comments added across five reviewers produced more comments in total than one instructor, which lead to more implemented revisions; (3) instructor suggestions for content additions were occasionally misunderstood by peers and led to a reduction in overall paper quality.

4.5 Future Research

This study is one of the first to examine at the validity of students' comments, so there are many directions to consider for future studies. The most obvious would be to address the limitations of this study. First, reviewers (i.e. peers versus instructors) should be randomly assigned to be able to measure the impact the comments have on writing performance. For example, solutions have been associated with more revision (Nelson & Schunn, 2009), so students, who offered more solutions than the content instructor, may help their peers to improve their paper more than the instructor. Also, a broader range of instructors and courses should be examined. For example, how teaching assistants' comments compare to other instructors and students should be examined since they are often the primary givers of feedback on writing. The current study only examined one content instructor and one writing instructor. We purposely selected

instructors that could be considered typical (e.g., instructors with considerable experience in teaching writing). Therefore, these instructors were expected to generalize to the typical WID content instructor and writing instructor. Of course, more junior instructors or instructors with little experience teaching writing may have different commenting styles.

Other important future studies will be necessary to determine which feedback features (i.e., praise, solutions, explanations) are indeed more useful. It is currently unclear whose feedback style (content instructor, writing instructor, or peer) is more helpful. Future research should also be done on the support needed by students in order to produce high quality feedback. We hypothesize that well designed rubrics and incentives for good reviews are likely to be important.

Finally, the impact of completing the reviewing task on writing performance should be considered. Wooley (2007) found that when students provide comments on their peers' writing prior to writing their own paper, their first draft is of better quality than those who reviewed their peers' writing after completing their first draft. Therefore, when an instructor is providing the feedback and grades, the students may be missing out on a potential learning opportunity.

4.6 Summary/Conclusion

Overall, students' comments seem to be fairly similar to instructors' comments, and therefore further validating the use of peer review. However, some differences were found. These differences between students and instructors and between content and writing instructors could help inform how to better prepare both instructors and students in providing feedback on writing. More research is necessary to determine which approach to feedback is best. If the student approach is indeed more helpful, then both writing instructors and content instructors may wish to include more praise and be more succinct. If the writing instructor approach is more helpful, then the content instructors and students should provide more solutions, specifically about high prose issues. If the content instructor approach is more helpful, then the writing instructors and students could point out more problems, specifically about content issues. These results are likely to rely on two key features of the SWoRD system, that is the specific rubrics provided to the students and the accountability system for accuracy and helpfulness. These results could generalize to any peer-review process that incorporates these key features.

References

Ashbaugh, H. A., Johnstone, K. M., & Warfield, T. D. (2002). Outcome assessment of a writing-skill improvement initiative: Results and methodological implications. *Issues in Accounting Education, 17*(2), 123-148. doi: 10.2308/iace.2002.17.2.123

Cho, K. (2005). *When multi-peers give better advice than an expert: The type and impact of feedback given by students and an expert on student writing.* Cho, Kwangsu: U Pittsburgh, US.

Cho, K., & Schunn, C. D. (2007). Scaffolded writing and rewriting 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. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Dragga, S. (1991). Responding to technical writing. *Technical Writing Teacher, 18*(3), 202-221.

Fife, J. M., & O'Neill, P. (2001). Moving beyond the written comment: Narrowing the gap between response practice and research. *College Composition and Communication, 53*(2), 300-321. doi: 10.2307/359079

Flower, L. (1986). Detection, diagnosis, and the strategies of revision. *College Composition and Communication, 37*(1), 16-55. doi: 10.2307/357381

Herrington, A. (1992). Composing one's self in a discipline: Students' and teachers' negotiations. In M. Secor & D. Charney (Eds.), *Constructing rhetorical education* (pp. 91-115). Carbondale, IL: Southern Illinois University Press.

Hillocks, G., Jr. (1984). What works in teaching composition: A meta-analysis of experimental treatment studies. *American Journal of Education, 93*(1), 133-170.doi: 10.1086/443789

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 31*(1), 159-174. doi: 10.2307/2529310

Loftus, G. R., & Masson, M. E. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review, 1*(4), 476-490. doi:10.3758/BF03210951

Nathan, M. J., & Koedinger, K. R. (2000). An investigation of teachers' beliefs of students' algebra development. *Cognition and Instruction, 18*(2), 209-237.doi:10.1207/S1532690XCI1802 03

Nelson, M. M., & Schunn, C. D. (2009). The nature of feedback: How different types of peer feedback affect writing performance. *Instructional Science,37(4)*. doi:10.1007/s11251-008-9053-x

Ochsner, R., & Fowler, J. (2004). Playing devil's advocate: Evaluating the literature of the wac/wid movement. *Review of Educational Research, 74*(2), 117-140. doi:10.3102/00346543074002117

Palincsar, A. S., & Brown, A. L. (1984). Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. *Cognition and Instruction, 1*(2), 117 - 175. doi:10.1207/s1532690xci0102 1

Prince, M. (2004). Does active learning work? A review of the research. *Journal of Engineering Education, 93*(3), 223-231. doi: 10.1002/j.2168-9830.2004.tb00809.x

Salahu-Din, D., Persky, H., & Miller, J. (2008). *The nation's report card: Writing 2007*: National Assessment of Educational Progress (NAEP).

Smith, S. (1997). The genre of the end comment + composition instruction: Conventions in teacher responses to student writing. *College Composition and Communication, 48*(2), 249-268. doi:10.2307/358669

Smith, S. (2003a). The role of technical expertise in engineering and writing teachers' evaluations of students' writing. *Written Communication, 20*(1), 37-80. doi:10.1177/0741088303253570

Smith, S. (2003b). What is "Good" Technical communication? A comparison of the standards of writing and engineering instructors. *Technical Communication Quarterly, 12*(1), 7 - 24. doi:10.1207/s15427625tcq1201 2

Sperling, M., & Freedman, S. (1987). A good girl writes like a good girl. *Written Communication, 4*(4), 343-369. doi: 10.1177/0741088387004004002

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

Wooley, R. S. (2007). *The effects of web-based peer review on student writing*. Unpublished Dissertation, Kent State University.

Appendix A

Reviewing Prompts

General Reviewing Guidelines

There are two very important parts to giving good feedback. First, give very *specific comments* rather than vague comments: Point to exact page numbers and paragraphs that were problematic; give examples of general problems that you found; be clear about what exactly the problem was; explain why it was a problem, etc. Second, *make your comments helpful*. The goal is not to punish the writer for making mistakes. Instead your goal is to help the writer improve his or her paper. You should point out problems where they occur. But don't stop there. Explain why they are problems and give some clear advice on how to fix the problems. Also keep your tone professional. No personal attacks. Everyone makes mistakes. Everyone can improve writing.

Prose Transparency

Did the writing flow smoothly so you could follow the main argument? This dimension is not about low level writing problems, like typos and simple grammar problems, unless those problems are so bad that it makes it hard to follow the argument. Instead this dimension is about whether you easily understood what each of the arguments was and the ordering of the points made sense to you. Can you find the main points? Are the transitions from one point to the next harsh, or do they transition naturally?

Your Comments:
First summarize what you perceived as the main points being made so that the writer can see whether the readers can follow the paper's arguments. Then make specific comments about what problems you had in understanding the arguments and following the flow across arguments. Be sure to give specific advice for how to fix the problems and praise-oriented advice for strength that made the writing good.

Your Rating:
Based on your comments above, how would you rate the prose of this paper?

☐	7. Excellent	All points are clearly made and very smoothly ordered.
☐	6. Very good	All but one point is clearly made and very smoothly ordered.
☐	5. Good	All but two or three points are clearly made and smoothly ordered. The few problems slowed down the reading, but it was still possible to understand the argument.
☐	4. Average	All but two or three points are clearly made and smoothly ordered. Some of the points were hard to find or understand.
☐	3. Poor	Many of the main points were hard to find, and/or the ordering of points was very strange and hard to follow.

☐	2. Very poor	Almost all of the main points were hard to find and/or very strangely ordered.
☐	1. Disastrous	It was impossible to understand what any of the main points were and/or there appeared to be a very random ordering of thoughts.

Logic of the Argument

This dimension is about the logic of the argument being made. Did the author just make some claims, or did the author provide some supporting arguments or evidence for those claims? Did the supporting arguments logically support the claims being made or were they irrelevant to the claim being made or contradictory to the claim being made? Did the author consider obvious counter-arguments, or were they ignored?

Your Comments:

Provide specific comments about the logic of the author's argument. If points were just made without support, describe which ones they were. If the support provided doesn't make logical sense, explain what that is. If some obvious counter-argument was not considered, explain what that counter-argument is. Then give potential fixes to these problems if you can think of any. This might involve suggesting that the author change their argument.

Your Rating:

Based on your comments above, how would you rate the logical arguments of this paper?

☐	7. Excellent	All arguments strongly supported and no logical flaws in the arguments.
☐	6. Very good	All but one argument strongly supported or one relatively minor logical flaw in the argument.
☐	5. Good	All but two or three arguments strongly supported or a few minor logical flaws in arguments.
☐	4. Average	Most arguments are well supported, but one or two points have major flaws in them or no support provided.
☐	3. Poor	A little support presented for many arguments, or several major flaws in the arguments.
☐	2. Very poor	Little support presented for most arguments, or obvious flaws in most arguments.
☐	1. Disastrous	No support presented for any arguments, or obvious flaws in all arguments presented.

Insight beyond core readings

This dimension concerns the extent to which new knowledge is introduced by a writer. 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?

Your Comments:

First summarize what you think the main insights were of this paper. What did you learn if anything? Listing this clearly will give the author clear feedback about the main point of writing a paper: to teach the reader something. If you think the main points were all taken from the readings or represent what everyone in the class would already know, then explain where you think those points were taken or what points would be obvious to everyone. Remember that not all points in the paper need to be novel, because some of the points need to be made just to support the main argument.

Your Rating:

Based on your comments above, how would you rate the insights of this paper?

☐	7. Excellent	I really learned several new things about the topic area, and it changed my point of view about that area.
☐	6. Very good	I learned at least one new, important thing about the topic area.
☐	5. Good	I learned something new about the topic area that most people wouldn't know, but I'm not sure it was really important for that topic area.
☐	4. Average	All the main points weren't taken directly from the class readings, but most people would have thought that on their own if they would have just taken a little time to think.
☐	3. Poor	Some of the main points were taken directly from the class readings; the others would be pretty obvious to most people in the class.
☐	2. Very poor	Most of the main points were taken directly from the class readings; the others would be pretty obvious to most people in the class.
☐	1. Disastrous	All the points stolen directly from the class readings.

Appendix B

Table B1. Coding Scheme categories and definitions

Category	Definition
Type of Feedback (kappa = .91)	
summary	A list of the topics, a description of the claims, or an identified action.
praise	A complimentary comment or identification of a positive feature.
problem/solution	Identifying what needs to be fixed and/or suggesting a way to fix an issue.
Type of Problem/Solution (kappa = .78)	
problem	Only a problem is explicitly identified.
solution	Only a solution is explicitly offered.
both	Both a problem and solution is provided.
Type of Affective Language (kappa = .66)	
compliment	An explicit compliment or positive modifier used to describe a criticism.
downplay	A reviewer minimizes the degree to which a problem is bad
questions	A question is used to identify a criticism or probe for more information
neutral	The language used to identify a criticism neutral or a matter-of-fact.
Localization of the Problem/Solution (kappa = .63)	
localized	A location was provided, so the issue can be easily found.
not localized	A location was not provided, so the issue cannot be easily found.
Scope of the Problem/Solution (kappa = .60)	
greater level	References a criticism that affects the whole paper.
midlevel	Criticism about a paragraph or part of paper (such as rough conclusion).
word-sentence	Criticism about a sentence or word (such as misspelled words).
Explanation of the Problem (kappa = .54)	
absent	Either none or a vacuous/circular explanation about a problem is provided.
content	Clarifying why a problem exists or how it is bad in a particular way.

Focus of the Problem (kappa = .69)	
assignment	Addresses problems with the assignment details.
combination	Addresses more than one focus.
high prose	Addresses problems that were defined in the rubric.
low prose	Addresses lower-level problems (such as grammar).
substance	Addresses problems with the content of the paper.
Explanation of the Solution (kappa = .58)	
absent	Either none or a vacuous/circular explanation about a solution is provided.
content	Clarifying why the solution fixes an issue or in what way it will be better.
Focus of the Solution (kappa = .58)	
assignment	Offers suggestions about assignment details.
combination	Offers suggestions about more than one focus.
high prose	Offers suggestions about the rubric.
low prose	Offers suggestions about lower-level problems (such as grammar).
substance	Offers suggestions about the content of the paper.

Appendix C

Table C1. Frequency and proportions by reviewer type (I: instructor; CI: content instructor; WI: writing instructor; P: peer; HP: high peer; LP: low peer; bold indicates significant differences)

				Reviewer-type				
Comment Type	**Total I**	**CI**	**WI**	**d CI vs.WI**	**Total P**	**HP**	**LP**	**d I vs. P**
word count	358	369	347	0.2	272	284	260	0.7
idea unit	6.9	7.4	6.4	0.4	7.8	8.1	7.4	-0.4
summary	0.12	0.14	0.11	0.4	0.10	0.08	0.12	0.3
praise	0.17	0.13	0.20	-0.4	0.31	0.31	0.31	-0.8
criticism	0.71	0.73	0.69	0.2	0.59	0.62	0.56	0.7
CRITICISM								
explicit problems	0.56	0.85	0.26	3.2	0.69	0.71	0.66	-0.7
explicit solutions	0.73	0.51	0.95	-2.6	0.62	0.60	0.64	0.6
problem & solution	0.29	0.35	0.23	0.6	0.31	0.32	0.30	-0.1
any mitigation	0.29	0.23	0.34	-0.5	0.29	0.29	0.28	0.0
compliment	0.11	0.06	0.17	-0.7	0.19	0.18	0.20	-0.5
downplay	0.00	0.01	0.00	0.3	0.03	0.05	0.00	-0.5
questions	0.17	0.16	0.18	-0.1	0.07	0.05	0.09	0.7
localization	0.56	0.61	0.51	0.5	0.49	0.53	0.45	0.4
greater	0.45	0.43	0.48	-0.2	0.46	0.42	0.49	0.0
midlevel	0.43	0.46	0.41	0.2	0.42	0.43	0.41	0.1
word-sentence	0.11	0.11	0.11	0.0	0.12	0.14	0.10	-0.1
PROBLEMS								
explanation	0.49	0.53	0.46	0.2	0.34	0.31	0.37	0.5
assignment	0.01	0.02	0.00	0.4	0.01	0.01	0.02	-0.1
combination	0.02	0.03	0.00	0.6	0.02	0.02	0.01	0.0
high prose	0.25	0.29	0.22	0.3	0.50	0.54	0.46	-1.0
low prose	0.12	0.08	0.17	-0.4	0.19	0.20	0.17	-0.3
substance	0.60	0.58	0.62	-0.1	0.28	0.23	0.33	1.1
SOLUTIONS								
explanation	0.42	0.34	0.49	-0.5	0.31	0.34	0.28	0.4
assignment	0.02	0.04	0.00	0.3	0.00	0.00	0.00	0.2
combination	0.04	0.03	0.04	-0.1	0.02	0.01	0.03	0.2
high prose	0.45	0.28	0.63	-1.5	0.43	0.45	0.42	0.1
low prose	0.06	0.09	0.03	0.5	0.26	0.30	0.22	-1.1
substance	0.41	0.52	0.30	0.8	0.28	0.24	0.32	0.5