Writing in natural sciences: Understanding the effects of different types of reviewers on the writing process

Melissa M. Patchan°+, Christian D. Schunn° and Russell J. Clark*

°LRDC, University of Pittsburgh | USA

+Department of Management, University of Antwerp | Belgium

*Physics & Astronomy, University of Pittsburgh | USA

Abstract: In undergraduate natural science courses, two types of evaluators are commonly used to assess student writing: graduate-student teaching assistants (TAs) or peers. The current study examines how well these approaches to evaluation support student writing. These differences between the two possible evaluators are likely to affect multiple aspects of the writing process: first draft quality, amount and types of feedback provided, amount and types of revisions, and final draft quality. Therefore, we examined how these aspects of the writing process were affected when undergraduate students wrote papers to be evaluated by a group of peers versus their TA. Several interesting results were found. First, the quality of the students' first draft was greater when they were writing for their peers than when writing for their TA. In terms of feedback, students provided longer comments, and they also focused more on the prose than the TAs. Finally, more revisions were made if the students received feedback from their peers—especially prose revisions. Despite all of the benefits seen with peers as evaluators, there was only a moderate difference in final draft quality. This result indicates that while peer-review is helpful, there continues to be a need for research regarding how to enhance the benefits.

Keywords: peer review; teaching assistant; audience effect; commenting style; revision

journal of WRITING RESEARCH

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

During the 1960s, academics began to see the importance of writing as pedagogy for enhancing learning, which developed into the Writing Across the Curriculum (WAC) and Writing in the Discipline (WID) movements (for a review, see Ochsner & Fowler, 2004). WAC aimed at students in their first two years of college with a focus on enhancing students' general knowledge. By contrast, WID aimed at students in their final two years and to refine their disciplinary knowledge—that is, the knowledge of how to read and write within their discipline. This disciplinary knowledge comprises subject matter knowledge as well as genre knowledge (Bazerman, 2008; Prior, 2006). The current study questions how well the typical WID environment supports its writing learning goals.

In a number of countries, including the US, two methods are frequently used for evaluating writing in undergraduate natural science courses: either a graduate student teaching assistant (TA) assigned to teach the course provides feedback on writing or a peer-review process is used in which the students grade and provide feedback to one another. We chose a context (i.e., large undergraduate introductory physics course) that typifies the common problem of incorporating writing in courses with high enrollment numbers and overloaded TAs. The contrasting methods for feedback in this setting are especially interesting because each method appears to emphasize different types of knowledge that are both necessary for writing. Across these methods, it is important to examine whether students are receiving sufficient writing instruction that balances the various core types of knowledge critical for writing success.

Several types of knowledge critical to effective writing have been identified—including discourse community knowledge, subject matter knowledge, genre knowledge, rhetorical knowledge, and writing process knowledge (Beaufort, 2004). Further, as Beaufort noted, each type of knowledge is likely to overlap with the other types of knowledge. Thus, there is no clear demarcation between what demonstrates subject matter knowledge and what demonstrates genre knowledge. That is, it is very hard to examine all writing moves or pieces of feedback on writing and categorize each clearly and uniquely in a single type of knowledge. However, the lack of clear lines of demarcation does not mean these distinctions are meaningless. Individuals receive training that focus on different types of knowledge, which produces different skill levels and possibly different values being placed on the relative importance of the types of knowledge.

For example, graduate students in physics would have developed considerable knowledge of physics content (e.g., which questions are important, how experiments are setup, how analyses are done) and some genre knowledge specific to writing in physics (e.g., from reading articles). In contrast, they would have relatively little rhetorical knowledge and knowledge about the writing process because they likely received little writing instruction and few opportunities to practice as writers. Further, they might have received feedback on writing that focused primarily in factual errors of the subject matter.

For many of natural science courses, the TAs are non-native English speaking graduate students. As a result, these differences in training are further amplified (Braine, 2002; Reinhardt, 2007). These TAs are likely to be even less proficient in their rhetorical knowledge and writing process knowledge, which is likely to reflect in their commenting practices.

Peers, on the other hand, are just beginning to develop content knowledge and may have only a shallow understanding of the concepts. In addition, they will have very little experience with the genre of writing in physics (e.g., examples of scientific writing, outlines of what is expected in each section). Instead, they will have had relatively more training about rhetorical situations and the writing process from their high school English classes and freshman composition courses. Furthermore, the feedback on their writing will likely have focused on the ways in which their arguments were expressed in addition to issues of factual correctness.

These differences between the two possible evaluators are likely to affect multiple aspects of the writing process. In this context, the writing process refers to all of the steps of a multi-draft writing assignment: writing a first draft, the evaluation of the first draft, and revising the first draft based on feedback received. First, the evaluator becomes the apparent audience of the paper and therefore could affect how the first draft is written. Second, the differences in prior experiences will likely influence what kinds of problems are detected, how they are valued, and thus what kinds of comments are provided. Third, the different kinds of comments could in turn affect the types of revisions made by the student and the resulting quality of the final draft. Therefore, how TAs and undergraduate students differentially impact each of these steps should be considered. The next three sections will address the expectations for each.

1.1 How do the two types of evaluators influence quality of the first draft?

Writing is supposed to be communicative; therefore, to whom the text is communicating (i.e., audience) should influence the writing process. Early writing research has shown that expert writers frequently consider the intended audience during the writing process in order to make decisions about which content to include and how to organize that content (Berkenkotter, 1981; Flower & Hayes, 1980). Interestingly, while novice writers have demonstrated the ability to adapt their texts for different audiences, the highest quality text was not always the one being evaluated for a grade.

Several researchers compared texts written for instructors (for a grade) to texts written for peers (not for a grade) and found that the texts written for the peers was more organized, contained richer content, and use clearer language; these texts were noticeably higher quality (Cohen & Riel, 1989; Gallini & Helman, 1995; Ward, 2009). Sato and Matsushima

(2006) compared texts written to be as accurate as possible to texts written so a reader could accurately recreate an abstract object. The participants who wrote texts for a reader spent more time planning and writing. In addition, the texts had shorter sentences, more subsidiary and meta descriptions, and were considered to be higher quality.

In each of the instances where the text was written for evaluation, the intended audience was someone with greater subject matter knowledge. This situation is common in written instruction; throughout elementary school, secondary school, and university, students frequently write only for their instructor, who has greater subject matter knowledge (Ede, 1979). However, this type of audience is especially awkward because it violates a major social norm. According to Gricean's Maxim of Quantity, "Do not make your contribution more informative than is required," a writer should not communicate information that the reader already knows (Grice, 1975, p. 45). In struggling with providing enough information to be considered accurate but also not providing information that the reader already knows, it may not be too surprising that the texts written for instructors were less coherent and did not contain as rich content.

In the current study, we sought to demonstrate that the lower quality texts are not resulting from the presence of evaluation, but rather the level of familiarity with the content being written about would affect the quality of the writing. Therefore, students were either evaluated by their TA or evaluated by a small group of peers. Because the TA had more subject matter knowledge and deeper knowledge of the particular assignment that was to be described in the paper than the peers, the first drafts that were written for peers were expected to be of higher quality than those written for their TA.

1.2 How do the two types of evaluators affect the type of feedback provided?

The varying levels of expertise for the two types of evaluators would likely affect the amount of feedback provided and on what that feedback focused. First, there would be a large difference in the workload for the evaluator. As the instructor, the TA would need to grade all of the students' papers. This amount is frequently as much as 20–25 papers, but sometimes can be well over 50 papers. In comparison, students are typically only assigned a small subset of the papers (e.g., 1 to 5 papers). Thinking about this large difference in workload, one may come to the conclusion that peers would be able to spend more time per paper than the instructor, which would result in more feedback. However, in prior research, undergraduate students either provided less feedback (Cho, Schunn, & Charney, 2006) or the same amount of feedback (Patchan, Charney, & Schunn, 2009) as instructors.

Several subject variables could explain this unexpected result. One important factor is how much the evaluator values providing feedback. There are some dedicated instructors who spend a lot of time per paper regardless of the number of papers to be evaluated, and there are some instructors who do not feel strongly about teaching writing that might spend a minimal amount of time on each paper. The same variability is likely to be seen with students as evaluators. Another important factor is experience with evaluation. Most instructors have evaluated many papers, which would allow them to provide feedback more efficiently (i.e., they could grade more papers in a shorter amount of time). By contrast, students have very little experience evaluating writing, so they would likely need more time per paper in their evaluations. Because of these additional variables, it is unclear which type of evaluator would be able to provide more feedback on a one-to-one comparison.

Despite the possible variability in one student review to one TA review comparisons, the number of evaluators for each paper would most likely affect how much feedback is provided. If an instructor was the only evaluator of student writing, then the students would receive feedback from only one source. In peer-review, a small group of students could be used to evaluate each paper. Therefore, the amount of feedback provided is multiplied by the size of the group. In previous pilot work, each student's review frequently commented on different problems (almost 75% of the comments are unique problems). Thus, there would be no question that the overall number of problems being detected by a group of peers would be greater than the amount detected by a single instructor. In addition, when multiple students comment on the same problem, students would be less able to rationalize that the problem was just a quirky personal preference ignoring and resulting in fewer problems being ignored.

Finally, as mentioned previously, the prior experiences of the non-native English speaking TAs are likely to result in strong content matter knowledge, moderate genre knowledge, and weak rhetorical and writing process knowledge. In contrast, the undergraduate students are likely to have weak content matter and genre knowledge and moderate rhetorical and writing process knowledge. These relative differences in knowledge will likely influence what kinds of problems in written documents are detected, how they are valued, and thus what kinds of comments are provided to students. Of course, there will be many issues that sit at the intersection of these types of knowledge and may be detected through multiple routes (e.g., as a violation of the genre and rhetorical issues); other problems may be detected with only basic levels of knowledge that are shared across the two different groups. Therefore, there might also be considerable overlap across problems detected and comments provided.

Several studies have specifically examined the focus of feedback provided by content instructors and undergraduate students. Feedback provided by engineering instructors focused primarily on the accuracy of the content in the text and sometimes the conventions of the disciplinary genre (Smith, 2003a, 2003b). While the undergraduates provided more solutions overall, the history instructor's solutions were more likely to focus on the subject matter issues (Patchan et al., 2009). This evidence would support the hypothesis that the focus of the feedback would depend on the types of expertise of the evaluator.

In the current study, the TAs were expected to focus less on rhetorical and writing process issues than the peers because the TAs would likely feel that it is not their job to do so. In addition, the TAs were expected to focus more on subject matter issues than the peers because the TAs would feel more confident in their diagnoses.

1.3 How do the two types of evaluators affect the revisions?

Many researchers have examined how writers of varying expertise revise their work, but in these contexts, the writer revised without the aid of feedback (e.g., Flower, Hayes, Carey, Schriver, & Stratman, 1986; Myhill & Jones, 2007). Without any guidance, students typically edit their papers rather than revise their text (e.g., correcting typos, misspelled words, grammatical errors, and maybe reword some sentences). These low-level revisions are not likely to result in significant draft quality differences. However, with very little training, students are able to change how they revise. After just eight minutes of training on how to make global revisions, students were able to make more global revisions and their papers' quality improved more so than students who were just asked to revise text without receiving the same training (Wallace & Hayes, 1991). This finding supports Fitzgerald's (1987) conclusion that feedback seemed to improve the quality of revisions.

As previously mentioned, instructors and peers are likely to comment in different ways. These differences are especially important because certain types of feedback seem to be more useful than others. Writers benefited the most from feedback that was general and focused on global issues (Cho & MacArthur, 2010; Strijbos, Narciss, & Dunnerbier, 2010). Furthermore, the accuracy of the criticisms and the justifications for the suggested revisions had an impact on the quality of the revisions being made (Gielen, Peeters, Dochy, Onghena, & Struyven, 2010). While these features of feedback seem to impact revision, no studies have examined how feedback focused on subject matter versus rhetorical issues affect the types of revisions.

Several studies have examined at a more general level whether the type of the evaluator affected students' revisions. Writers revised more successfully after reflecting on a particular audience (Jucks, Schulte-Lobbert, & Bromme, 2007; Midgette, Haria, & MacArthur, 2008). While students were more likely to use feedback provided by instructors, their revisions were more successful when implementing peer feedback (Miao, Badger, & Zhen, 2006).

In the current study, the types of revisions were expected to reflect the types of feedback that was provided in a straightforward fashion: if the comments focused on the content, then the students are likely to make changes regarding the content, and if comments focused on the prose, then the students are likely to make changes regarding the prose. Because the TAs were expected to comment more on the content, students who received feedback from their TA were expected to make more content revisions. Because the peers were expected to comment more on the prose, students who received feedback from their peers were expected to make more prose changes.

1.4 Hypotheses

In the current study, we examined the relative efficacy of two common, pragmatic choices for providing feedback on writing (i.e., non-native English TAs or peer-review) by looking at the influences of the two choices on the writing process. In an introductory physics lab course, students wrote lab reports to be graded either by their non-native English TA or a group of their peers. Several expectations about each of the steps of a multi-draft writing assignment have been identified regarding the general difference between groups:

1. The quality of the first draft was expected to be higher when writing for their peers than when writing for their TA.
2. Students were expected to focus more on prose issues than TAs.
3. TAs were expected to focus more on physics issues than students.
4. Students, who received feedback from their peers, were expected to make more prose changes.
5. Students, who received feedback from their TA, were expected to make more physics changes.
6. As a result of these expectations, students who initially wrote to their peers were expected to produce final drafts of higher quality than those who initially wrote to their TA.

There are likely to be differences across individuals within each group, but these individual differences are not the focus of this paper.

2. Method

2.1 Overview

In this study, we experimentally manipulated the type of evaluators of students' writing in order to draw some conclusions about the effects of different audiences on the writing process. More specifically, we examined the writing quality of first and second drafts of two reports, the feedback provided by students and TAs, and the changes made between first and second drafts.

2.2 General Method

Course context & participants. The participants in this experiment included 211 students who were enrolled in an Introduction to Laboratory Physics course and 11 TAs who taught the lab sessions. The course included a diverse student population, with many different majors represented (e.g., biology, chemistry, nursing, pharmacy, and pre-medicine). Almost 60% of the students were female, and less than 3% were not native-English speakers. Most of the students were sophomores (33%), juniors (43%), or seniors (17%). While none of the TAs were native English speakers (primarily Cantonese or Mandarin native speakers), all of the TAs were sufficiently fluid in written and spoken English to meet university guidelines for TAs responsible for face-to-face instruction.

The course was designed to expose students to the experimental process in physics by obtaining, analyzing, and presenting their own experimental results. One form of presenting results involved writing formal lab reports, which included five sections: abstract, introduction and theory, experimental setup, data analysis, and conclusion (see Appendix A for details). Each student was required to write two formal lab reports: one on ballistic motion and one on the charge to mass ratio (e/m) of the electron. For each report, the first draft was graded and commented upon by either the TA or four peers (see Appendix B for reviewing guidelines). Then the students were expected to revise their report based on the feedback they received. The TA graded all of the final drafts. The first draft, reviewing activities, and the final draft of both reports were collectively worth 30% of their final grade.

Design. Each lab section was randomly assigned to one of two conditions: peer-first and TA-first. In the peer-first condition, four peers were assigned to review each of the first reports using an online peer-review system. In the TA-first condition, the students' TA would review their paper. The source of feedback was then counterbalanced for the second report (i.e., the TA reviewed the second report in the peer-first condition, and four peers reviewed the second report in the TA-first condition).

Review support structures. Students had one week to write their first draft, and they were required to submit them electronically. Where they submitted their papers depended on who would be reviewing them. Therefore, students were aware of who their audience was. Students, whose papers were to be reviewed by other peers, submitted their first draft through the SWoRD (Scaffolded Writing and Rewriting in the Discipline) system, an anonymous web-based reciprocal peer-review system. Students, whose papers were to be reviewed by their TA, turned their papers in through the university's *Blackboard* system. While there were minor interface differences in how students submitted their papers to their peers or TA, both interfaces were basically a simple document-upload interface. Thus this difference would be highly unlikely to affect the quality of the paper, which would be completed by the time the document upload step began.

The SWoRD system was used to facilitate the process of distributing documents and reviews to peers (Cho & Schunn, 2007), while the more straightforward distribution of documents to and feedback from TAs was handled by Blackboard. Different systems were used for both generalization and pragmatic considerations. In terms of generalization, multiple peer review requires some kind of method to support that process (whether face-to-face or online), and TA feedback would typically involve a different method. Pragmatically, SWoRD did not have a facility for distributing all papers to TAs and Blackboard did not have a facility for distributing papers to a fixed number of peers. Critical to the internal validity of the experiment, instructions for the rating dimensions and providing comments were exactly the same between peers and TAs. Further, the basic process was the same between the two interfaces: authors upload papers, reviewers download papers, reviewers upload comments, authors view comments, authors revise documents, and authors upload final drafts.

After the first draft deadline, the students and TAs had one week to review the papers assigned to them. The TAs reviewed all of the papers in their lab (ranging from 11 to 18 students), and each peer reviewed four papers. Both students and TAs provided feedback and a rating from a 7-point scale on three different dimensions; they received detailed instructions about how to review papers. Two of the dimensions (i.e., Introduction, Theory and Experimental Setup; Abstract, Conclusion, and References) focused on the prose aspects of the report, and one of the dimensions (i.e., Data Analysis and Results) focused on the physics aspect of the report.

For the reports to be reviewed by students, the SWoRD system automatically distributed each paper to four peers. The pool of students used to randomly assign reviewers was the whole class (i.e., all 211 students), not just students from the same lab. Therefore, it would be nearly impossible for the students to identify the author or reviewer unless the student included it (which they were instructed not to).

After receiving their feedback, the students then revised their reports using that feedback. The TAs graded their students' final drafts for both reports. They did not have access to the peers' comments, so the peer evaluations would not influence the final quality marks.

Data Analysis. We had several expectations about each of the steps of a multi-draft writing assignment have been identified. In order to test all of our hypotheses, several different samples of the data were necessary. Therefore, each step of the writing assignment will be discussed as if it was its own study.

3. Study 1A: Initial Draft Quality

First, we wanted to know how the two types of evaluators influenced the quality of the first draft. That is, do peers write differently when the apparent audience is their peers versus their TA? The quality of the first draft was expected to be higher when writing for their peers than when writing for their TA. To eliminate bias in ratings, expert ratings were used to compare the quality of the students' first draft.

3.1 Method

Sample. Of the 211 students, a sample of 50 students, sufficiently large for statistical analysis, was randomly selected. Half of the students were from the peer-first condition and half from the TA-first condition. In order to use a representative sample, the number of students selected from each section was based on how many students there were in that section.

Coding. The ratings of two outside experts (i.e., a writing expert and a content expert) were used instead of the students and TAs ratings to obtain fully comparable and face-valid evaluations. The writing expert was a full-time faculty member from the university writing center, and the content expert was a native English physics graduate student who had prior experience with teaching this particular lab. These experts were specifically chosen for high face-validity, which should reduce the likelihood that important aspects of the writing would be missed. In addition, both experts were given the same rubric as the students to guide their attention to criticism factors of writing in these documents. The two experts rated the quality of the 50 students' first and second drafts of both reports—a total of 200 papers. The writing expert rated the first and third dimension because these focused more on prose aspects of the lab report, and the writing expert would thus provide the highest validity judgments for those dimensions. No content knowledge was necessary to evaluate the prose in these sections. The content expert rated the second dimension, which focused on the domain knowledge of the report, and thus a content expert was necessary for evaluation of that dimension. Experts were blind to condition and draft. Correlations between first and second drafts (by dimension) indicated that the experts were able to produce reliable ratings (writing expert: $r(164) = .86$, $p < .001$; content expert $r(94) = .64$, $p < .001$).

3.2 Results & Discussion

The dependent measure of interest was the quality of the first draft, which was measured by combining the expert ratings into one composite score (out of 21 possible points). The quality was then compared across the two types of audience (i.e., peers versus non-native English TAs). Students wrote better papers for their peers than they did for their TA (see Figure 1), $t(83) = 2.24$, $p = .03$.

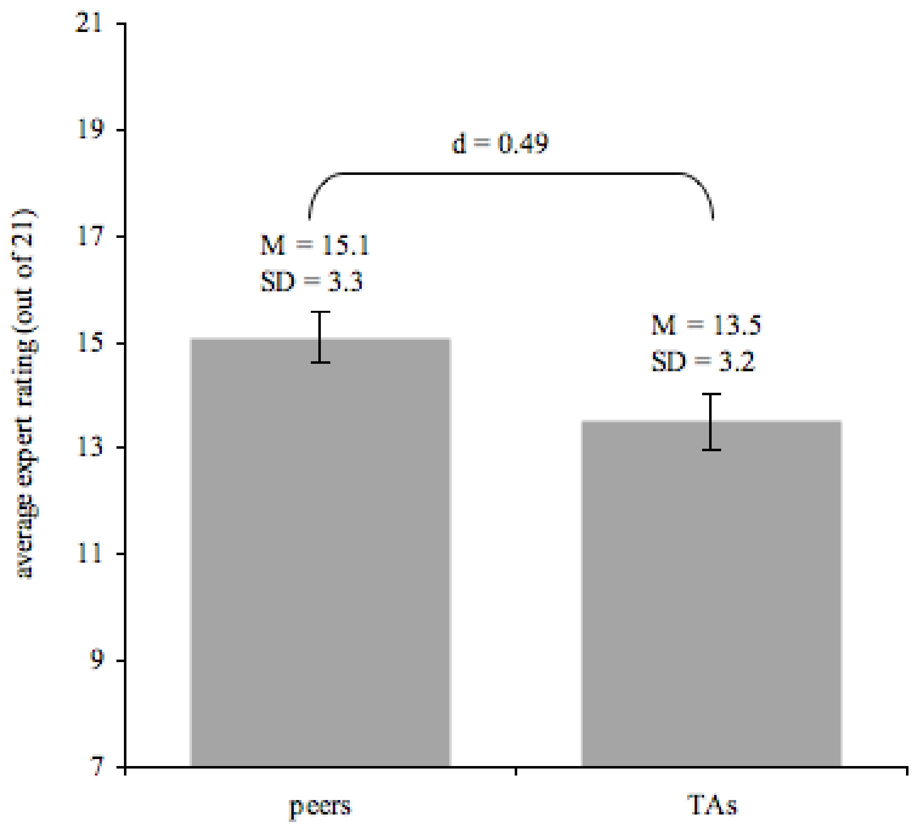

Figure 1: Average total score of first draft by initial evaluator (with SE bars)

One may wonder whether this audience effect generalizes across all reviewing dimensions, especially since the focus of the dimensions vary from prose features to physics features. The expert ratings for each dimension were analyzed as within-subjects variables. None of the interactions involving dimension were significant: the first draft ratings for all three dimensions were higher when the audience was their peers rather than their TA (see Table 1).

Table 1: Means and standard deviations of first draft by reviewing dimension for peer evaluators and TA evaluators

	Peers	>	TAs	
	M	*SD*	*M*	*SD*
Introduction, Theory, and Experimental Setup	5.3	1.3	4.7	1.6
Data Analysis and Results	5.0	1.8	4.3	1.6
Abstract, Conclusion, and References	5.0	1.1	4.6	1.4

4. Study 1B: Feedback Features

Next we wanted to know how feedback differed when provided by undergraduate students versus non-native English TAs. Students were expected to focus more on prose issues than TAs. TAs were expected to focus more on physics issues than students. Feedback provided by students and TAs was segmented and analyzed in detail.

4.1 Method

Sample. A total of 29 peers were sampled from the seven sections in which the first report was peer-reviewed, and all of the five non-native English TAs who commented on first drafts for the first report were included. All of the feedback from these 34 reviewers was first broken down by dimension (i.e., Introduction, Theory and Experimental Setup; Abstract, Conclusion, and References, Data Analysis and Results). Because feedback could focus on many different issues, all of the feedback was segmented using idea unit as the unit of analysis (i.e., the feedback about poor transitions was separated from feedback about having a clear thesis). Each piece of feedback could have anywhere from 3 to 37 segments. A total of 796 feedback segments were analyzed.

Coding. The comments of interest were the criticism comments (i.e., feedback that describes a problem and/or solution rather than praise or summary statements). Therefore, the comments were first categorized as criticism or not (102 segments were double-coded with 100% agreement). Each piece of criticism was then categorized as either focusing on physics (e.g., calculations, data, equations, interpretations, theory) or focusing on prose (e.g., audience, content, flow, length). Some of the criticism comments were too vague to be labeled as physics or prose, so they were just labeled as nonspecific. For example, it was unclear whether a comment like “This section needs a lot of work!” is referring to physics or the prose aspect of the paper. Less than 5% of the segments were labeled nonspecific. The reliability of coding was checked by two research assistants who independently code a subset of the segments; 102 segments were double-coded with moderate reliability (Kappa

= .74), well above the typical threshold of .41 for acceptable reliability (Landis & Koch, 1977).

4.2 Results & Discussion

The first dependent variable of interest was length (i.e., number of words in all of the comments provided by one reviewer). The length of the comments was examined across the two types of reviewers (i.e., peers vs. non-native English TAs). Analyses compared the average peer (single review) to the average TA (single review). Individual peers provided longer reviews than the TAs (see Figure 2), $t(32) = 2.43$, $p < .001$. This result replicated for each reviewing dimension. Further, taken from a writer's perspective, the four peers assigned to provide feedback would provide comments almost ten times longer than a TA's comments.

The second dependent variable of interest, orthogonal to length of comment, was type of comments (i.e., focusing on physics or focusing on prose). This variable was also compared across the two types of reviewers (i.e., the average peer compared to the average TA). Overall, individual peers comment on marginally more issues than individual TAs (see Figure 3), $t(32) = 1.98$, $p = .06$. As expected, peers focused more on prose issues than the TAs, $t(32) = 2.43$, $p = .02$. Interestingly, peers focused on physics issues just as frequently as the TAs, $t(32) = 1.28$, $p = .21$. However, from a writer's perspective, peers would comment on almost seven times as many ideas overall than the TA—11 times as many prose issues and six times as many physics issues.

5. Study 1C: Revisions

Next, we wanted to know how revisions differed when students received different amounts and types of feedback. Students, who received feedback from their peers, were expected to make more prose changes. Students, who received feedback from their TA, were expected to make more physics changes. To test this prediction, the revisions students made between their first and second drafts were analyzed.

5.1 Method

Sample. The same sample of 50 students used to test the audience effect was also used to examine the effect of feedback on revisions. A list of students' revisions was created using the compare documents function in Microsoft Word 2003. On average students made 32 changes between the first draft and second draft. However, some students did not make any changes, and some students made over 100 changes (even up to 195 changes). A total of 3,165 revisions were coded.

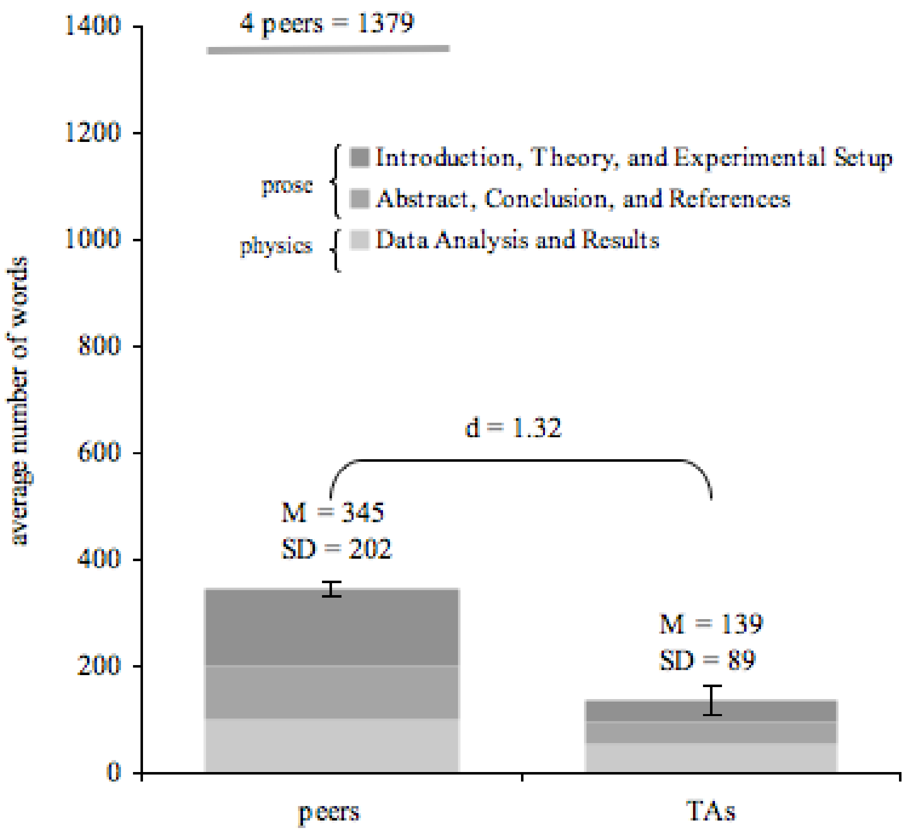

Figure 2: Average number of words per review by reviewing dimension and by type of evaluator (with SE bars)

Coding. The revisions were first categorized as either focusing primarily on physics or prose issues. Some revision types were obviously physics (e.g., changing data reported), and some revisions were obviously prose (e.g., changing verb tense). However, some revision types, such as changing wording, could be labeled either as physics or prose. To determine which label was most appropriate, the rule "could someone without any physics knowledge make this change?" was used to label all revisions. All of the prose revisions were further categorized as either low prose or high prose. This determination was based on length of the change. Revisions that involved more than one sentence were considered to be a high prose revision, and revisions that were contained within a sentence were considered to be a low prose revision. The reliability of coding was checked by two research assistants who independently code a subset of the segments. For the type of change (i.e., physics or prose), 202 revisions were double-coded with a resulting moderate reliability (Kappa = .55). For type of prose change (i.e., low prose or high prose), 85 prose revisions (found in the set of 202 double-coded revisions) were double-coded, also with a resulting moderate reliability (Kappa = .71).

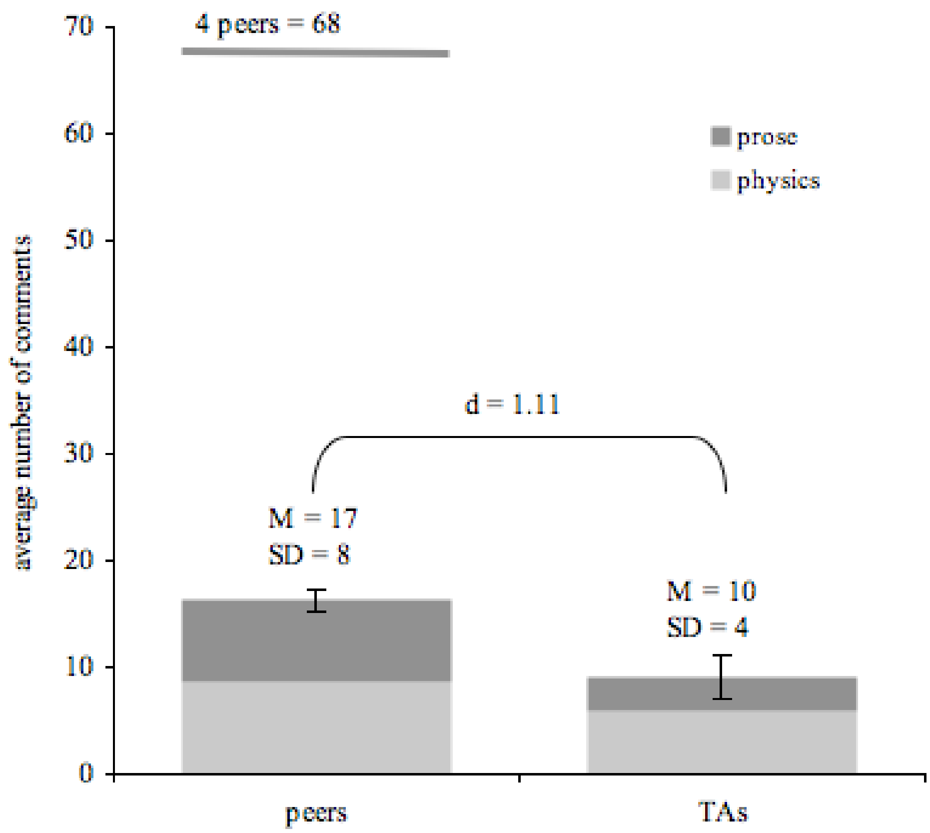

Figure 3: Average number of comments by comment type and by type of evaluator (with SE bars)

5.2 Results & Discussion

The dependent variable of interest is the average number of revisions made by the students. Again, this variable was compared across the two types of reviewers (i.e., peers vs. non-native English TAs). Overall, students made more changes when feedback came from peers than when feedback came from TAs, $t(81) = 3.82$, $p < .001$.

There were several types of changes one could make (i.e., a change focused on physics issues, a changes focused on high-level prose issues, or a change focused on low-level prose issues). The number of physics changes did not differ by who provided feedback, but students who received feedback from peers made more prose changes than those who received feedback from TAs (see Figure 4), $t(81) = 4.67$, $p < .001$. This result replicated for both low-level changes and high-level changes. Overall more low-level changes were made than high-level changes, $t(162) = 7.96$, $p < .001$.

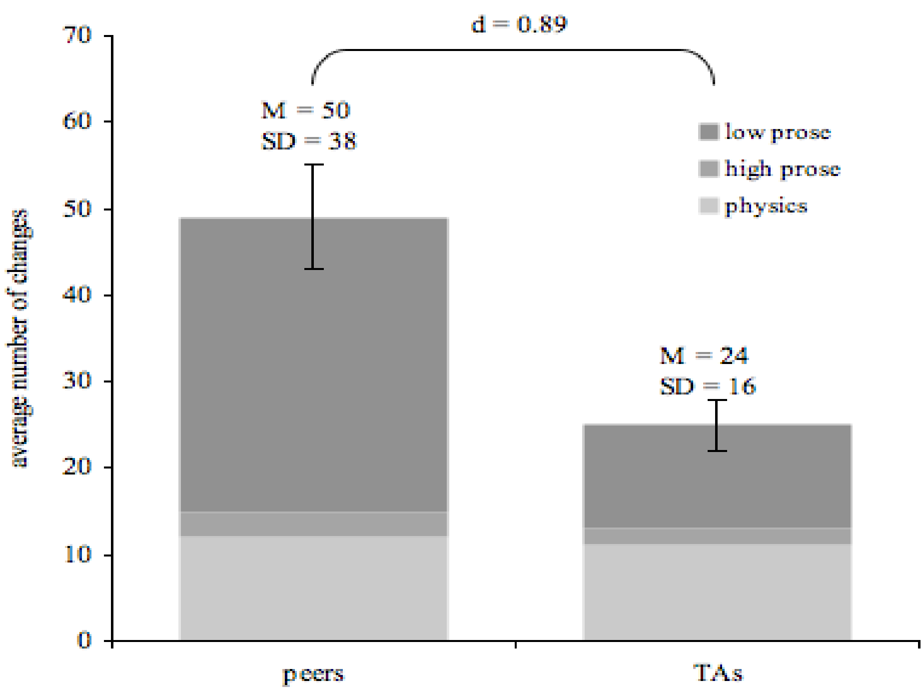

Figure 4: Average number of revisions by type of revisions and by type of evaluator

6. Study 1D: Final Draft Quality

Finally, we wanted to know whether the differences in amount and type of revisions resulted in differences in the final draft quality. Students were expected to produce the best quality writing when initially writing for their peers because they were expected to start with better first drafts, get more feedback from their peers, and make more revisions. Again, expert ratings were used to compare the quality of the students' final draft. The same sample and ratings obtained for Study 1A was used in for this analysis.

6.1 Results & Discussion

Because there were differences across conditions in first draft quality, comments received, and revisions made, there were likely to be differences in the final draft quality. Indeed, students who initially wrote for their peers produce final drafts that were significantly higher in quality than those who initially wrote for their TA, $t(83) = 1.95$, $p = .05$ (see Figure 5), although here the effect is only moderate.

7. General Discussion

7.1 Summary of Results

First, there was a significant audience effect: the quality of students' first draft was greater when they were writing for their peers than when they were writing for their TA. This audience effect generalized across all reviewing dimensions (i.e., when focusing on prose or physics issues).

In regards to commenting, students provided longer comments than TAs. They also discussed more ideas. While not completely surprising, this finding was not what was predicted. In previous research, instructors' comments were longer than students' comments (Cho et al., 2006; Patchan et al., 2009). One reason for the difference could be that the non-native English graduate students in the current study may have found the task of commenting more difficult due to their relatively weaker language proficiency and therefore commented less.

From a writer's perspective, the difference between the length of a peer's review and a TA's review was amplified because students receive feedback from four of their peers. Therefore, on average a student would receive almost 1400 words of feedback from their peers but only a little more than 100 words from their TA (as indicated in Figure 2). These large differences could be quite important. Due to the large amount of feedback provided by peers, the writer should have a number of high quality comments to help with revision, even though not all students are experts in writing or providing feedback. On the other hand, the students would need to sift through a lot more feedback to determine which

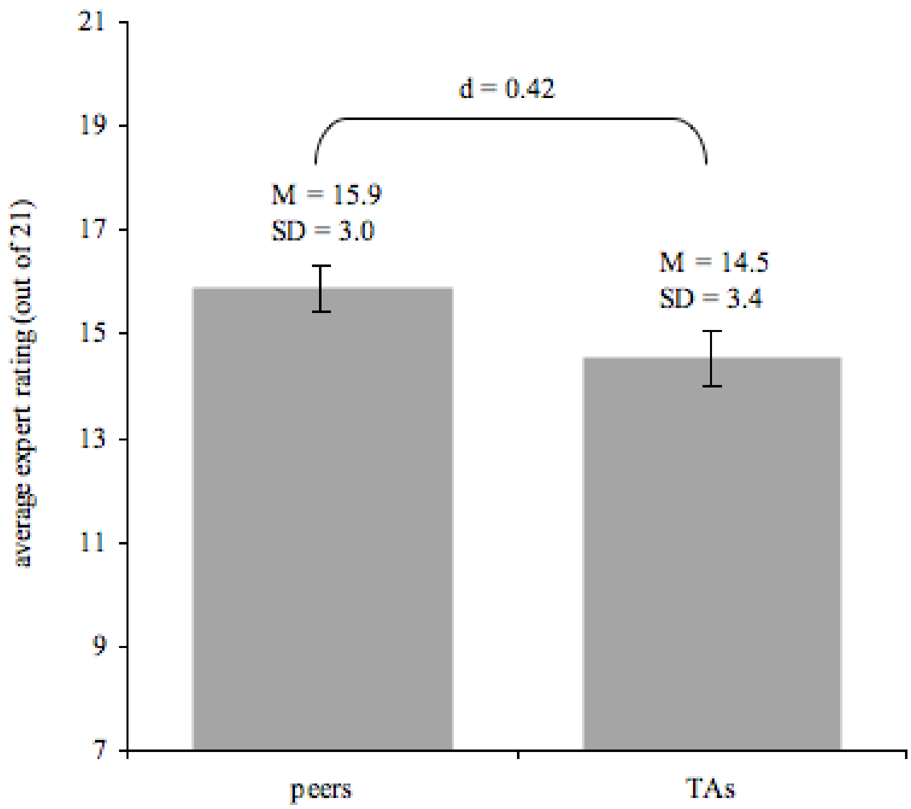

Figure 5: Average total score of final draft by initial evaluator (with SE bars)

comments were the most appropriate or useful. Students could become overwhelmed with the need to sift through so much feedback, but the sifting-through-feedback task could also be beneficial to the student. That is, it could help the student better understand the writing process (i.e., what works and what does not) and the reviewing process (i.e., what makes for good feedback).

While there were no differences overall in the amount of physics versus prose comments, students commented about prose more than the TAs. Surprisingly, there were no differences between students and TAs on the amount of physics comments. The TAs were expected to comment more about physics problems than the students because this area would fall within their expertise. Perhaps difficulties with English prevented the TAs from being able to create comments focusing on physics.

Several differences in revisions occurred, which likely resulted from the differences in comments received. Overall, students made more changes after receiving comments from their peers, which is not surprising because they received more feedback from their peers. In addition, more prose changes were made based on feedback from peers rather than in response to the TA. More specifically, students made a lot more low-level prose revisions after receiving feedback from peers rather than their TA. On the other hand, there were no differences in the number of physics changes. This finding partially supports the expectation that revisions would correspond with the content of the feedback. Students provided more prose changes than the TAs, so more prose revisions would be expected when feedback came from peers. However, the group of peers provided about six times as much physics feedback as the TA alone.

There were several possible explanations for this surprising null effect on the number of physics revisions. First, there may have been redundancy among peer comments. Therefore, the number of unique physics suggestions provided by the group of peers may have been equivalent to the number provided by the TA. This explanation seems unlikely. In previous studies, we have examined how often students comment on the same issues and found that a majority of student comments in a set of four or five peer comments on a paper involve unique issues (i.e., each peer finds additional errors not noted by other peers). Another explanation could be that the TAs would be considered the authority on content issues, and thus students would likely implement all of the TAs suggestions but only some of the peer comments (given some skepticism of their peers' ability to comment on the physics content). Therefore, they may be more selective in which comments they actually implement.

Even though there were differences across conditions in first draft quality, comments received, and revisions made, there was only a moderate difference on final draft quality. Why was the condition effect not larger for second drafts? One possibility is that the shift from peer grader to TA grader undercut the forcefulness of peer grades (i.e., comments could be more safely ignored when the evaluator changed). However, this explanation would have predicted fewer revisions rather than more revisions in the peer feedback condition. A more likely explanation involves differences in the quality of feedback—the TAs' comments may have been better than the peers' comments. Covariate analyses, if all comments to each author had been coded, would provide some data to this point. In addition, comparing the quality of comments becomes further complicated when considering the students' ability to comprehend the comments. Experts could rate whether the comments, if acted upon, could improve the content of the paper. In doing so, we may see that the TAs appeared to provide higher quality comments than the students. However, it would be unclear whether the students who read the comments would understand them enough to be able to successfully revise their paper. For example, Cho and colleagues (2006) found that instructors thought other instructors comments were more accurate and useful than peer comments while peers saw instructor and peer comments as equally accurate and useful. This complex issue should be examined in future research.

7.2 Implications

The current study provides additional support for the validity of using peer-review to evaluate student writing. Most studies that examine the validity of peer-review focus on the validity of the ratings (Cho et al., 2006; Cho, Schunn, & Wilson, 2006). Few studies have compared the comments provided by students to comments that would have been provided by an instructor (Cho et al., 2006; Patchan et al., 2009). While these studies have shown that students are able to provide comments that are similar to instructors, many individuals (both instructors and students) are still skeptical about the effectiveness of peer comments. The results of the current study show that the peer-review process can have benefits on a multitude of levels. First, students who write for their peers create first drafts of better quality than when writing for an instructor. Second, students can receive more feedback about their draft from their peers than they would from their instructor—especially more feedback focusing on how to improve the prose. Finally, when students receive feedback from their peers, they are likely to make more revisions—again, these changes tend to focus on the prose. While there seem to be a lot of benefits to the students, the final draft quality is not that much different when initially written for peers. At the very least, this result does support the use of peer-review as acceptable practice, in that it could produce outcomes for student authors that would be, at minimum, equivalent to what they would get if an instructor evaluated their work.

We are not advocating for replacing TA feedback entirely with peer feedback. Faculty, TAs, and students are likely to find that situation very controversial. Instead, we are arguing that peer feedback may be just as effective as TA feedback for first drafts. TA workload (which is often quite high in lab contexts with significant writing components) may be reduced, or shifted to other feedback tasks by having peers provide feedback on first drafts. Alternatively, peer feedback may be added to TA feedback on first drafts, perhaps providing an optimal balance of audiences and prose/content knowledge oriented feedback. We also note that peers are likely to be obtaining some benefits from providing feedback to peers (Wooley, 2007; Wooley, Was, Schunn, & Dalton, 2008), although we did not examine such effects in this study.

7.3 Caveats

There are several caveats to these finding that must be taken under consideration. First, the focus of the current study was to examine whether students are receiving sufficient writing instruction in natural science contexts that balances the various core types of knowledge (i.e., subject matter knowledge, rhetorical knowledge, writing process knowledge) critical for writing success. This distinction is very important because the two most common evaluators of student writing are peers and graduate students. In many physics courses offered at universities in the US, the graduate students teaching the courses with the high writing components (i.e., labs) tend to be non-native English speaking. This language distinction is most prominent in physics courses and not as strong in other natural science disciplines, such as biology. Therefore, our results regarding the effects of TAs on student writing may only generalize to settings in which the TA is a non-native English speaker. On the other hand, the physics context is very similar to other natural science disciplines in that the instructors have a much deeper understanding of the domain than the peers. If this content knowledge difference is what drives the current findings, rather than the language proficiency difference, then the results of the current study may generalize to all natural science disciplines.

A second caveat involves the nature of the peers. Although coming from a variety of disciplinary backgrounds, one might wonder how the TA versus peer results would have changed if the students had all been significantly stronger writers (e.g., at more selective universities/colleges than the currently studied setting, or in honors/advanced sections), or significantly worse writers (e.g., as occurs at less selective universities/colleges than the currently studied setting). A related dimension involves student motivation levels, which has several elements: 1) do the students perceive grades in the given course as important and thus worth significant effort overall, and 2) do the students more specifically perceive the peer review task as important and thus worth significant effort? Future research should examine these variables more carefully.

The third caveat to be considered is the structure of the reviewing task in the current study. Very detailed guidelines were provided to both the students and the TAs about how to review a paper. These guidelines not only offered structure to the type of feedback that should be provided, but the guidelines also set specific expectations for what should be considered important in writing a formal laboratory report. Previous research has shown that just a little guidance could greatly affect one's writing behaviors (Wallace & Hayes, 1991). Therefore, these results may only generalize to those instances where specific guidelines for reviewing are provided. Peers are quite likely to be influenced by reviewing guidelines. If students are only told to point out problems with their peer's writing, they are likely to just edit the paper. Instead of using broad instructions, students are providing specific prompts to help them focus on important prose and content issues. For example, one comment prompt requested reviewers to "describe any problems in understanding how the final results relate to the theoretical predictions." Once prompted, students were capable of formulating comments regarding more sophisticated issues that would otherwise have been overlooked. TAs are also very influenced by reviewing guidelines (Smith Taylor, 2007), but perhaps in a different way. Interestingly given the same prompts, the TAs from our study did not provide as much prose comments as the peers. For example, one comment prompt asked "Was the writing appropriate for the target audience (your fellow classmates)? If not, then explain why." Despite these specific prompts, the TAs continued to focus more on the content issues. From our study, we do not know how the reviews from each group would compare when 'unstructured', although prior research would suggest that both peer and TA comment quality would significantly suffer.

7.4 Future Research

One possible direction for future research could be to compare the effects of non-native English graduate student TAs versus native English graduate student TAs, again looking separately at the different steps in the writing-commenting-revision process. Would the students perceive these two groups as different audiences? How does the feedback provided by the TAs differ? The results could help determine whether the current study's findings generalize to all graduate-student TAs. It would also be useful to know whether these results are specific to science writing or could they generalize to other domains.

7.5 Summary/Conclusion

In the current study, students wrote better first drafts when they expected their peers to grade that draft rather than their TA. Students provided longer comments and more comments about prose problems than the TAs. As a result, students made more revisions after receiving feedback from peers than from their TA. Interestingly, these differences did not seem to have a large impact on the final draft score, indicating that students may need additional training on how to provide higher quality feedback. Based on these results, peer feedback should be used to supplement (not replace) the TA feedback—peers can focus on the writing aspects of the paper, while the TA focuses on the physics/domain aspect of the paper.

References

Bazerman, C. (2008, November). *US Writing Across the Curriculum and Writing in the Disciplines in Context.* Paper presented at the Proceedings of the International Conference « de la France au Québec : l'Ecriture dans tous ses états », Poitiers, France.

Beaufort, A. (2004). Developmental Gains of a History Major: A Case for Building a Theory of Disciplinary Writing Expertise. *Research in the Teaching of English, 39*(2), 136-185.

Berkenkotter, C. (1981). Understanding a Writer's Awareness of Audience. *College Composition and Communication, 32*(4), 388-399.

Braine, G. (2002). Academic literacy and the nonnative speaker graduate student. *Journal of English for Academic Purposes, 1*(1), 59-68.

Cho, K., & MacArthur, C. (2010). Student revision with peer and expert reviewing. *Learning and Instruction, 20*(4), 328-338.

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.

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.

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.

Cohen, M., & Riel, M. (1989). The effect of distant audiences on students' writing. *American Educational Research Journal, 26*(2), 143-159.

Ede, L. S. (1979). On Audience and Composition. *College Composition and Communication, 30*(3), 291-295.

Fitzgerald, J. (1987). Research on Revision in Writing. *Review of Educational Research, 57*(4), 481-506.

Flower, L., & Hayes, J. R. (1980). The Cognition of Discovery: Defining a Rhetorical Problem. *31*(1), 21-32.

Flower, L., Hayes, J. R., Carey, L., Schriver, K., & Stratman, J. (1986). Detection, Diagnosis, and the Strategies of Revision. *College Composition and Communication, 37*(1), 16-55.

Gallini, J. K., & Helman, N. (1995). Audience awareness in technology-mediated environments. *Journal of Educational Computing Research, 13*(3), 245-261.

Gielen, S., Peeters, E., Dochy, F., Onghena, P., & Struyven, K. (2010). Improving the effectiveness of peer feedback for learning. *Learning and Instruction, 20*(4), 304-315.

Grice, H. P. (1975). Logic and conversation. In R. Cole & J. Morgan (Eds.), *Syntax and semantics: Vol.3, Speech acts* (pp. 41-58). New York: Academic Press.

Jucks, R., Schulte-Lobbert, P., & Bromme, R. (2007). Supporting experts' written knowledge communication through reflective prompts on the use of specialist concepts. *Zeitschrift Fur Psychologie [Journal of Psychology], 215*(4), 237-247.

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

Miao, Y., Badger, R., & Zhen, Y. (2006). A comparative study of peer and teacher feedback in a Chinese EFL writing class. *Journal of Second Language Writing, 15*(3), 179-200.

Midgette, E., Haria, P., & MacArthur, C. (2008). The effects of content and audience awareness goals for revision on the persuasive essays of fifth- and eighth-grade students. *Reading and Writing, 21*(1-2), 131-151.

Myhill, D., & Jones, S. (2007). More than just error correction - Students' perspectives on their revision processes during writing. *Written Communication, 24*(4), 323-343.

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.

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.

Prior, P. (2006). A sociocultural theory of writing. *Handbook of writing research*, 54-66.

Reinhardt, J. (2007). *Directives usage by ITAs: An applied learner corpus analysis.* Unpublished Doctoral Dissertation, The Pennsylvania State University.

Sato, K., & Matsushima, K. (2006). Effects of audience awareness on procedural text writing. *Psychological Reports, 99*(1), 51-73.

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

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.

Smith Taylor, S. (2007). Comments on lab reports by mechanical engineering teaching assistants: Typical practices and effects of using a grading rubric. *Journal of Business and Technical Communication, 21*(4), 402-424.

Strijbos, J. W., Narciss, S., & Dunnebier, K. (2010). Peer feedback content and sender's competence level in academic writing revision tasks: Are they critical for feedback perceptions and efficiency? *Learning and Instruction, 20*(4), 291-303.

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

Ward, M. (2009). Squaring the Learning Circle Cross-Classroom Collaborations and the Impact of Audience on Student Outcomes in Professional Writing. *Journal of Business and Technical Communication, 23*(1), 61-82.

Wooley, R. S. (2007). *The Effects of Web-Based Peer Review on Student Writing.* Unpublished Empirical, Kent State University.

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

Appendix A

Formal Lab Report Contents

1. Abstract - This is a one paragraph complete summary of the experiment and your results.

2. Introduction and Theory - Here you should describe the basic physical theory behind the experiment and how the experiment tests the theory. Include all the relevant formulas and how they are used. Write in your own words; do not copy text from the lab manual. You are allowed to use figures from the lab manual provided you reference them in the report

3. Experimental Setup - Explain the experimental procedure and how the data was collected. Your report should include enough information that anyone reading it could reproduce your experiment. Be specific about the equipment used and the experimental procedure. If necessary, include labeled diagrams of the equipment. You may copy and reference figures from the manual, but do not copy the written material.

4. Data, Analysis and Questions - Include your data in tables with labels and captions. Likewise, show the results of your analysis in tables and plots with labels and captions. Your results here should support your conclusions in the next section. You should also include the answers to all questions asked in the lab manual.

5. Conclusion - The conclusion should be a summary of the whole experiment but it should focus on the final results. Include your own observations and comments regarding the experiment and suggest ways to improve the results. Explain how your data and analysis support the theory behind the experiment. If they do not support the theory, then explain why not.

6. References – A list of references you used in writing the report.

Appendix B

Formal Lab Report Review Form

When you review, there are two very important parts to giving good feedback.

1. 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.

2. Make your comments helpful. The goal is not to punish the author for making mistakes. Instead your goal is to help the author 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.

This evaluation form is divided into three parts:

1. Introduction, Theory and Experimental Setup
2. Data Analysis and Results
3. Abstract, Conclusion and References

For each part you will provide written comments for the author and a numerical score from 1 to 7, where 7 means excellent. There are guides below for how to assign the numerical values, but ultimately you should use your best judgment.

If you are using the SWoRD system (peer reviews) note that this form DOES NOT automatically submit your reviews to the system. After you are done, you need to access SWoRD, enter your numerical scores and copy and paste your review comments into the online form.

Introduction, Theory and Experimental Setup

Did the writing flow smoothly so you could understand the theory as it was tested with the experimental setup? 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 understand the text. Instead this dimension is about whether the author adequately explained the theory and the experimental setup used to test the theory. Is the purpose of the experiment clear? Is the experimental procedure described in sufficient detail? Are the transitions from one section to the next harsh, or do they transition naturally?

Your Comments

First summarize what you perceived as the purpose of the experiment and what it was testing so that the author can see whether the readers understood the writing in this section.

1. Describe any problems to understanding the theory as presented by the author.
2. Was the writing appropriate for the target audience (your fellow classmates)? If not, then explain why.
3. Describe any problems in understanding how the experimental setup tested the theory and avoided any likely experimental uncertainties.
4. If you thought some aspect of this section was done well, mention what was good about it. Be sure to give specific advice for how to fix any problems and praise-oriented advice for strengths that made the writing good.

Your Rating

1.	Excellent	The written explanation of the theory and the experimental method is clear and very easy to understand.
2.	Very Good	The written explanation of the theory and the experimental method is clear and somewhat easy to understand.
3.	Good	The written explanation of the theory and the experimental method could be made a little clearer.
4.	Average	The written explanation of the theory and the experimental method could be made a lot clearer.
5.	Poor	It is somewhat difficult to understand either the theory or the experimental method used from the writing.
6.	Very Poor	It is very difficult to understand either the theory or the experimental method used from the writing.
7.	Disastrous	Cannot understand either the theory or the experimental method used from the writing.

Data Analysis and Results

This dimension is about how the theoretical predictions compare to the experimental results. Are the theoretical predictions accurate; in other words did the author apply the theory to the experimental setup correctly so as to produce an accurate prediction? Did the author collect a good sample of data and was the analysis of the data performed correctly? Finally, did the final results agree with the theory within the experimental uncertainties and if not, did the author provide an adequate explanation for the discrepancy?

Your Comments

1. Describe any problems in understanding how the theoretical predictions were calculated.
2. Describe any problems in understanding how the data analysis calculations were performed.
3. Describe any problems in understanding how the final results relate to the theoretical predictions.
4. Describe whether significant deviations from the theoretical prediction (if any) are explained.
5. If you thought some aspect of the Data Analysis and Results section was well done then mention what was good about it.

Your Rating

1.	Excellent	The calculations are clear and complete. The results are consistent with theory and any unusual deviations are clearly explained.
2.	Very Good	The calculations are clear and complete. The results are reasonably consistent with theory and any unusual deviations are clearly explained.
3.	Good	The calculations are clear and complete. The results are somewhat consistent with theory and any unusual deviations are adequately explained.
4.	Average	The calculations are not clear or complete. The results are somewhat consistent with theory and any unusual deviations are adequately explained.
5.	Poor	Calculations are incomplete or unclear. The results are somewhat inconsistent with the theory and only a poor explanation for the disagreement is given.
6.	Very Poor	Calculations are incomplete or unclear. The results are inconsistent with the theory and only a poor explanation for the disagreement is given.
7.	Disastrous	Calculations are incomplete and unclear. The results are wildly inconsistent with the theory and no adequate explanation for the disagreement is given.

Abstract, Conclusion and References

This dimension concerns the content of the abstract and the conclusion. The abstract should be a short (one paragraph) summary of the whole report including the final results. The conclusion should also be a brief summary of the whole report but it should focus more on the final results and on the comparison of those results to the theoretical predictions. The conclusion may be longer than one paragraph.

Your Comments

1. Describe any problems with the abstract or conclusion length.
2. Describe any main ideas or results missing from the abstract.

3. Describe whether the conclusion focuses on the main results and any problems in explaining how the results relate to the theory.
4. Were references properly cited for materials (ideas, figures or text) that were included from external sources?
5. If you thought some aspect of the abstract or conclusion was well done, mention what was good about it.

Your Rating

1.	Excellent	The abstract briefly captures the main ideas of the whole report including the results. The conclusion briefly summarizes the whole report but focuses on the main results and how they relate to the theory. References were properly cited for all materials included from outside sources.
2.	Very Good	The abstract captures the main ideas of the whole report including the results but is a little too long. The conclusion briefly summarizes the whole report and focuses somewhat on the main results and how they relate to the theory. References were properly cited for all materials included from outside sources.
3.	Good	The abstract captures most of the main ideas of the whole report including the results but is a little too long. The conclusion briefly summarizes the whole report and only adequately focuses on the main results and how they relate to the theory. References were properly cited for all materials included from outside sources.
4.	Average	The abstract captures some of the main ideas of the whole report including the results but is too long. The conclusion briefly summarizes the whole report but does not adequately focus on the main results and how they relate to the theory. Some references were omitted for materials included from outside sources.
5.	Poor	The abstract misses some of the main ideas of the whole report and is a little too long. The conclusion summarizes most of the report and does not adequately discuss the main results and how they relate to the theory. Some references were omitted for materials included from outside sources.
6.	Very Poor	The abstract misses most of the main ideas of the whole report and is too long. The conclusion summarizes only some of the report and does not adequately discuss the main results and how they relate to the theory. Many references were omitted for materials included from outside sources.
7.	Disastrous	The abstract misses the main ideas of the whole report and is much too long. The conclusion fails to summarize the majority of the report and fails to discuss the main results and how they relate to the theory. No references were provided for materials included from outside sources.