Journal of the Scholarship of Teaching and Learning, Vol. 2, No. 1, February 2008, pp. 29 - 44.

Development of Student Writing in Biochemistry Using Calibrated Peer Review

Yasha Hartberg, Adalet Baris Gunersel, Nancy J. Simspon, and Valerie Balester[1]

Abstract: This study investigating the effectiveness of Calibrated Peer Review (CPR) ™ in a senior-level biochemistry class had three purposes: to (a) compare the CPR process for feedback with TA-generated feedback in improving students' ability to write scientific abstracts; (b) compare CPR results for males and females; and (c) observe whether CPR improved the quality of student writing. Statistical analyses of three assignments by 50 students indicated significant differences between CPR and TA feedback on student writing quality. In addition, while scores of students who received TA feedback decreased, scores of students who had CPR improved. Students also progressed in CPR-generated measures of their writing and reviewing abilities. A separate analysis including 256 students found no significant differences between males and females. In addition, students' writing showed statistically significant improvement in CPR-generated scores.

Keywords: Calibrated Peer Review, writing skills, peer response, peer critique, abstract, teaching assistants, computer-related gender differences.

Those who research and teach composition have long dealt with the relationships between quality of writing and quality of thinking, form and content, conceptual understanding and written expression. As colleges and universities increase attention given to improving writing competency by requiring writing-intensive courses in all disciplines, instructors of courses other than English composition are grappling with helping students learn to write. Efforts to improve student writing inevitably result in *more* student writing and, in turn, more responding to and grading of writing. In many cases, faculty rely on graduate teaching assistants (TAs) to grade and give feedback on student writing. While this can ease the time burden for faculty, reliance on TAs is not without its own challenges. Faculty need to teach their TAs how to recognize the degree to which student papers meet expectations and how to provide instructive feedback to students. Further, in many disciplines there are other aspects of instruction, such as facilitating laboratories or recitations, for which TAs are needed. Thus, college faculty teaching writing-intensive courses continually seek methods that make most effective use of time—their own as well as that of their students and their TAs.

An innovative educational tool—Calibrated Peer Review (CPR)™ —offers one approach to meeting these challenges. CPR is a Web-based program that was developed at UCLA for the Molecular Science Project, one of the NSF-supported Chemistry Systematic Reform Initiatives (http://cpr.molsci.ucla.edu/). CPR was developed to give students practice in both writing and peer review, since these are common processes for scientific research (Russell, 2001). After

[1] Biochemistry and Biophysics, Building room 210, Texas A&M University, College Station, TX 77843-2128, yasha@tamu.edu; Center for Teaching Excellence, 533 Blocker, MS 4246, Texas A&M University, College Station, TX 77843, bgunersel@tamu.edu and n-simpson@tamu.edu; Texas A&M University, 5000 TAMU, College Station, TX 77843-5000, v-balester@tamu.edu.

submitting their papers, students practice reviewing sample papers using an instructor-designed rubric, receive feedback on their reviews, and then critique their peers' work anonymously; each student's paper is graded by three peers. Students also assess their own work using the same rubric. (For a more detailed explanation of the steps of the CPR process, see Appendix 1.)

This study investigates the effectiveness of CPR in a senior-level biochemistry class. For the instructor of this course, student writing had long been a priority. At the time he heard about CPR, he had tried several approaches to teaching his TAs to grade student papers, including well-developed grading rubrics, but was dissatisfied with the results. Even with extensive training, TAs would inevitably slip into grading to standards that were different from those established by the rubrics, even when the TAs had substantial input into the rubric design. Therefore, in spite of worries about the amount of time that CPR would demand of his students, the instructor decided to try using this tool. After adopting CPR, he noticed improvement in student writing, and, although students did complain about the amount of time required, he also heard from some that CPR was helping them learn the material better. His sense of positive results motivated him to continue using CPR. However, he wanted a more systematic way of investigating CPR's effectiveness. He met with two faculty developers and the executive director of the student writing center, all of whom had also worked with CPR and posed similar questions about the tool. This joint curiosity led to the current study. Quantitative analysis of student scores, along with the instructor's input of his own experience, were used to address the following three questions: (1) Is the CPR process for evaluation and feedback at least as effective as feedback generated by TAs in improving students' ability to write scientific abstracts? (2) Are CPR results different for males and females? (3) Does CPR improve student writing of abstracts in a senior-level biochemistry course? For the course in this study, abstracts described the backgrounds, methods, results, and conclusions of a lab exercise performed in class.

I. Calibrated Peer Review (CPR).

CPR is built upon two pedagogical practices—writing and peer review—that are well supported by educational research. The Writing-Across-the-Curriculum movement has been broadly supported by institutions of higher education since 1985 (Barnett and Blumner, 1999). Studies indicate that writing not only aids the learning process, but also promotes the development of critical thinking skills (Klein, 1999; Paul, 1995; Sternberg, 1994). Well-crafted writing assignments promote active reading and critical thinking by having students use course concepts to confront problems, gather and analyze data, prepare hypotheses, and formulate arguments (Lowman, 1996; Wright, Herteis, and Abernethy, 2001). Writing helps students extend their knowledge, formulate new understandings, and structure rudimentary ideas into greater coherence (Herrington, 1997; Rivard, Stanley, and Straw, 2000). Finally, writing helps prepare students for future careers by helping them to "become better acquainted with the forms of writing required by various academic disciplines and professions" (Klein, 1999, pp. 203-204).

Research also points to the value of giving students opportunities to practice and guiding them in reviewing each other's work (Pope, 2005). Studies have found that peer review is an effective way of teaching and learning (e.g., Boud, 1990; Cutler and Price, 1995; Dochy, Segers, and Sluijman, 1999; Orsmond, Merry, and Callaghan, 2004; Pope, 2005; Reese-Durham, 2005; Sobral, 1997; Topping, 1998). For example, Orsmond et al. (2004) found that peer review gave students practice in developing criteria regarding performance and identifying the gaps between the actual and desired performance. Other studies have found that it leads to an increase in student performance on assessments, as well as an increase in the quality of learning output (Cutler and Price, 1995; Freeman, 1995; Reese-Durham, 2005; Stefani, 1992; Topping, 1998). It encourages students to be reflective (Boud, 1990) and may lead to a positive perception of peers (Topping, 1998) and greater satisfaction in their own productivity (Cutler and Price, 1995).

Research findings on student response to peer review are mixed. Reese-Durham's (2005) students reported that the feedback from their peers was constructive, clear, and helpful and that the process made them realize that they had to practice and improve their reviewing skills. Other studies indicate that students think peer review forces them to think and learn more (Falchikov, 1995; Wen and Tsai, 2006), lets them compare different approaches in writing and standards of work, and allows them to exchange information and ideas (Williams, 1992). In addition, peer review gives students the opportunity to learn the class content more effectively and to understand the assignment content and assessment process (Brindley and Scoffield, 1998). On the other hand, other researchers report that students have difficulty in criticizing friends and perceive grades given by peers to be arbitrary (Williams, 1992), worry about variations in how criteria are interpreted, distrust peers' evaluation abilities, and believe that assessment is the role of the instructor and not the student (Brindley and Scoffield, 1998).

The research cited above gives evidence that the design of CPR is pedagogically sound. The body of research specific to CPR is small, but positive. Instructor-reported experiences and a limited number of studies suggest that it is a tool that can help students master content, improve writing skills, and become more competent reviewers (Furman and Robinson, 2003; McCarty et al., 2005; Russell, 2001). Gerdeman, Russell, and Worden (2007) examined the development of 1330 students' writing and reviewing skills in an introductory biology course and found that students showed improvement in writing and reviewing over three CPR assignments. Margerum et al.'s (2007) survey with first-semester general chemistry students found that students felt they were becoming better technical reviewers with CPR assignments and that students mastered the class material through the reviewing process. Palaez's (2002) study compared the impact of peer review in CPR and the impact of traditional instruction on undergraduate nonscience majors' performance on physiology tests. After comparing test results of students who had used CPR and who had received traditional instruction, Palaez (2002) found that the performance of students who used CPR was equal to or better than the performance of those who received traditional instruction. The current study contributes to this body of research by using quantitative analysis of student scores, interpreted in the context of the instructor's experience.

II. Context for the Study.

While designated "senior-level," the biochemistry course was the first exposure most students had to biochemistry lab practices. The majority of the students enrolled were juniors and seniors. Students conducted laboratory experiments, wrote associated lab reports, and also wrote formal abstracts for a subset of the experiments. For the instructor, the abstract-writing assignment was important and was therefore weighted almost as heavily as the lab reports in determining course grades. The abstracts followed a strict, one-paragraph format consisting of a descriptive title, background information, objectives of the study, methods used, results generated and conclusions drawn. As an aid to students, the course lab manual contained an extensive discussion of abstracts including the purpose and function of an abstract in scientific writing, a description of each section of an abstract, and a detailed critique and revision of a student abstract. The students were also provided a tutorial that described strategies for writing abstracts.

Prior to 2005, student writing was graded by graduate teaching assistants. In 2005, the instructor introduced CPR as the process for having students write abstracts and receive feedback and grades on their papers. Implementation of CPR was not without its difficulties. Consistent with the literature cited earlier, some students resisted grading, and being graded by, their peers. However, the instructor also noticed that student writing was improving. To test the accuracy of the instructor's observations, this study compares abstracts written by students who used CPR with abstracts written by students whose papers were graded by TAs.

In 2004, students completed four writing assignments that were graded by teaching assistants. In 2005, CPR was introduced and students completed three assignments. The instructor decided to have students write fewer assignments in order to compensate for the fact that CPR requires more work than writing without the reviewing process. For both 2004 and 2005 classes, the writing assignments required students to complete a set number of related biochemical techniques and write an abstract describing purpose, methods and results.

III. Methods.

A. Participants.

For the comparison of TA feedback and CPR (research question 1), 50 students (22 male and 28 female) were selected at random, 25 from Fall 2004 (semester with TAs) and 25 from Fall 2005 (semester with CPR). For analysis of gender differences with CPR and CPR's effectiveness (research questions 2 and 3), all 256 students who used CPR in 2005 were included (71 male, 185 female). Detailed information on participants in different analyses is provided in the data analysis section.

B. Scoring Abstracts.

In order to establish an independent standard by which to evaluate student writing, a primary trait grading rubric was developed for abstract writing (Appendix 2). Primary trait scoring is well-suited to drawing attention to the rhetorical traits of a specific type of document, in this case a scientific abstract, most valued by a disciplinary practitioner (Lloyd-Jones, 1977; Odell, 1992).The course instructor selected the traits and their weight based on his methods of instruction, his directions to students, and his concept of an ideal abstract. With careful rater training, primary trait scoring can be a reliable means of judging what particular aspects of a writing task are being mastered. For example, primary trait scoring can show whether students in the sections using CPR are doing better on one trait than another.

Seven independent evaluators were selected from graduate students in biochemistry, genetics or toxicology, all of whom had demonstrated an ability to write in the scientific discipline. To minimize bias, evaluators who had no previous experience with the laboratory class were chosen.

To ensure that evaluators were only considering the quality of the text, all abstracts were formatted to give a uniform appearance. Any personal identifying information was removed and each abstract was given a code consisting of a word or an abbreviation designating the primary topic of the assignment followed by a randomly generated 4-digit number. The abstracts within each topic were arranged in numerical order, effectively randomizing the pool of abstracts.

Before grading the papers selected for this study, the evaluators were trained to be consistent. After reading over the rubric, the evaluators discussed the various criteria and asked the instructor questions if they had any. Each of the evaluators then scored a sample abstract according to the rubric. Scores were compared and differences were discussed with the course instructor, after which graders were given an opportunity to rescore the abstract. To ensure that the same standards were being maintained as grading progressed, this process was repeated several times with other sample abstracts until a reasonable consensus emerged.

Following the training, each of 150 abstracts (50 students, 3 abstracts for each) was scored according to the rubric by two evaluators (not including the instructor). On the rubric, a total of 50 possible points were available; the total score for each abstract was calculated by adding the scores of two evaluators. When the point difference between the two scores was greater than seven, a third grader scored the abstract. Then the final score was calculated by adding the two closest scores. On one occasion, a third score fell directly between the original scores in which case the two highest scores were added. The average difference between the two scores that were finally used to assess the abstract was 3.57; inter-rater reliability (Cronbach's Alpha) calculated using these pairs of scores was 0.887.

C. Data Analysis.

Research Question 1. In order to determine whether the CPR process for evaluation and feedback was at least as effective as feedback generated by TAs in improving students’ ability to write scientific abstracts, two analyses were conducted. The first two assignments completed in 2004 were identical to the first two assignments completed in 2005; thus, the first analysis included these assignments. First, a repeated measures analysis was conducted with the selected 50 students and a total of 100 abstracts. The within-subject factor was time (two assignments) and the between-subject factor was semester (CPR or TA). The dependent variable was the final score given by the independent evaluators.

The second analysis compared abstracts identified as high quality by TAs with those identified as high quality by peers through the CPR process. The purpose was to determine whether abstracts that were rated highly by either means would also be rated as high quality by the instructor. Sixteen abstracts that had been scored highly were selected, eight from the 2004 semester which had been scored highly by TAs and eight from the 2005 semester which had been scored highly by peers through the CPR process. Scores of the abstracts from 2004 were higher than 90 on a scale of 1-100, while text rating scores from the abstracts from 2005 were higher than 8.55 on a scale of 1-10. The abstracts were coded and randomized so that the instructor would not know which papers had been originally evaluated by TAs and which had been evaluated through CPR. The instructor then graded the abstracts with the same rubric used by the independent graders.

Research Question 2. In order to determine whether CPR results were different for males and females, the 256 students in all of the sections that used CPR in 2005 (71 males and 185 females) were included. A repeated measures analysis on three assignments completed with CPR was conducted. The dependent variables included six scores generated by CPR: overall grade, text rating, reviewer competency index, review score, self-assessment, and calibration score. (For explanations of each of these variables, see Appendix 3.) The within-subjects factor was assignment number and the between-subject factor was gender.

Research Question 3. To determine if student writing improved with the use of CPR, the 256 students who had taken the 2005 course with CPR were included in the analysis. An ANOVA was conducted. The independent variable was time (3 assignments), and the dependent variables included several scores generated by CPR: text rating, percent correct style, percent correct content, reviewer competency index, calibration deviation, and review deviation.

IV. Results.

A. Research Question 1.

When students from both semesters were considered as a group, there was no significant difference between the means on assignment one and the means on assignment two (*df*= 1, *F*= 0.053, n^2= 0.001, *p*< 0.819). However, there was a significant difference between results obtained with feedback from TAs and CPR (semester by time interaction) at alpha level 0.05 (*df*= 1, *F*= 5.880, n^2= 0.109, *p*=< 0.20). While students' scores improved in the semester with CPR over two assignments, scores declined in the semester with the TAs. (See Table 1 for descriptive statistics.)

Table 1. Descriptive Statistics.

		M	*SD*	*N*
Assignment 1	TA	29.9400	7.22599	25
	CPR	26.8000	7.78353	25
	Total	28.3700	7.60022	50
Assignment 2	TA	27.6600	6.25620	25
	CPR	29.5600	6.26274	25
	Total	28.6100	7.88132	50

The second analysis also bore interesting results. Among the selected high quality abstracts, the instructor scored abstracts written through the CPR process higher than the abstracts that had been graded by TAs on every rubric category except for categories 4 and 5 (Table 2). Category 4, which refers to background information and clarification of objectives, was scored higher for TA abstracts than the CPR ones. Scores for category 5, which refers to methods, were equal for TA abstracts and CPR ones.

B. Research Question 2.

Results indicate that there were no significant differences between the performance of males and females on CPR (assignment number by gender interaction) in any of the different scores (overall grade, text rating, review competency index, review score, self-assessment, and calibration score) (Table 3). This lack of difference suggests that CPR does not disadvantage students based on gender.

Table 2. Descriptive statistics.

Rubric Question	*Semester*	*M*	*SD*
1	TA	0.7500	0.70711
	CPR	1.7500	0.70711
2	TA	0.8750	0.99103
	CPR	1.0000	0.92582
3	TA	0.8750	0.99103
	CPR	1.1250	0.99103
4	TA	1.1250	0.64087
	CPR	0.7500	0.70711
5	TA	0.6250	0.74402
	CPR	0.6250	0.74402
6	TA	1.0000	0.92582
	CPR	1.1250	0.64087
7	TA	0.7500	0.88641
	CPR	1.3750	0.91613

Table 3. ANOVA Table.

	Df	*η2*	*F*	*P*
Overall Grade	2	0.001	0.358	0.699
Text Rating	2	0.004	0.956	0.385
RCI	2	0.003	0.825	0.439
Review Score	2	0.001	0.127	0.880
Self-Assessment	2	0.004	0.886	0.413
Calibration Score	2	0.011	2.769	0.064

C. Research Question 3.

In the ANOVA, all the variables (the different CPR-generated scores) showed statistically significant improvement. There were statistically significant increases in text rating (*df*= 2*, F*= 8.143*, p*< 0.000)*,* percent correct for style (*df*= 2*, F*= 39.709*, p*< 0.000), percent correct for content (*df*= 2*, F*= 20.700*, p*< 0.000), RCI (*df*= 2*, F*= 63.926*, p*< 0.000) and statistically significant decreases in calibration deviation (*df*= 39.918*, F*= 48.826*, p*< 0.000) and review deviation (*df*= 2,*F*= 9.4223*, p*< 0.000) (Table 4). The decrease in the deviation scores is desirable, as it suggests that students are internalizing the instructor's criteria for writing and are reaching a consensus about what constitutes effective writing.

V. Conclusions.

Results suggest that the CPR process for providing evaluation and feedback is more effective than TA-generated feedback in improving students' ability to write scientific abstracts. Over the course of two assignments, the quality of abstracts written under the guidance of TA-generated feedback decreased. This surprising result might reflect the difficulty of transmitting learning objectives through third parties. Despite careful efforts to ensure that TAs understood the instructor's expectations for writing abstracts, TAs might have an inherent tendency to form

Table 4. Descriptive Statistics.

	Assignment No.	*Mean*	*SD*
Text Rating	1	5.7443	1.88588
	2	6.2358	1.54098
	3	6.2747	1.58806
Percent Correct Style	1	69.8623	17.98941
	2	77.9921	13.32459
	3	82.3933	17.27221
Percent Correct Content	1	70.4885	16.21321
	2	77.1123	13.72686
	3	78.5122	15.74931
RCI	1	3.0451	1.52891
	2	3.9575	1.60006
	3	4.5953	1.60312
Calibration Deviation	1	1.8021	0.95548
	2	1.4724	0.87247
	3	1.0519	0.76723
Review Deviation	1	1.3677	0.88763
	2	1.2546	0.71577
	3	1.0779	0.65828

their own opinions about what constitutes a “good” abstract and, through their feedback, push students in a direction contrary to what the instructor had intended. From this perspective, CPR represents a more direct line of communication between instructor and student. Even though students evaluate each other’s work with relatively little direct supervision from the instructor, CPR requires that students repeatedly revisit the instructor’s expectations through the application of the instructor-generated grading rubric used in the calibration, peer review, and self-assessment stages of each assignment. This conscious engagement with those points the instructor had identified as being important could account for the improvement in the quality of student abstracts when CPR was used as the evaluation tool. It would also explain why CPR did a better job than TAs at identifying abstracts that match the instructor’s expectations as indicated by the blind test in which the instructor scored abstracts evaluated using CPR more highly than those that had been evaluated by TAs. This is consistent with other research that shows that the processes of understanding the instructor’s rubric and using it to review peers’ written work enhance the learning of critical content (e.g., Margerum, et al., 2007).

While, overall, students who used CPR wrote better abstracts than students who received TA feedback on their writing, the researchers wanted to know if there were any aspects of writing scientific abstracts that CPR did not address as effectively as TA generated feedback. The detailed analysis of the scoring rubrics suggested that TA generated feedback outperformed CPR in only one category, background and objectives. In the instructor’s experience, the background section of an abstract is particularly difficult for students to write if for no other reason than students have limited experience in the field. Evaluating backgrounds generally requires a certain breadth of knowledge in the discipline as well as some level of experience reading scientific literature. Students’ naïveté tends to restrict their ability to place what they have done in the laboratory into a broader scientific context, an essential function of the background section. It seems reasonable that this inexperience would also make it difficult for students to evaluate this part of the abstract in a peer review setting. As such, it would make sense that students would receive more useful feedback from TAs than from their peers in this category.

It is interesting to note that detailed analysis of the traits on the grading rubrics revealed only one aspect of writing scientific abstracts, the methods category, was equivalent between the two semesters. One might expect that TAs, who are usually more technically proficient than their students, would also provide more effective feedback on the technical details of the methods section. However, that was not the case for the student abstracts in this analysis. Despite their relative lack of experience, students apparently are as competent as TAs to review each other's methods.

Because a concern for female equity in computer-related fields started in the 1990s and was expected to continue into the new millennium (Bunderson and Christensen, 1995; Camp, 1997; Davies and Camp, 2000; Young, 2000), the researchers wanted to determine if female performance in CPR differed from male performance. Results indicated that there were no significant differences between the performance of males and females in the semester with CPR, which suggests that CPR does not disadvantage female students and that female students' competencies with the CPR software are similar to the competencies of male students. While some studies found gender differences in computer-related competence (e.g., Durndell and Thomson, 1997; Janssen Reinen and Plomp, 1997; Volman, 1997), this study is supported by various studies that found there were no differences between women and men in computer performance (e.g., Doornekamp, 1993; Fitzgerald, 1987).

One benefit of using CPR is that the program returns a wealth of data at the end of each assignment on virtually every aspect of student performance. This provides insight into student learning that is generally inaccessible to the instructor through more conventional assignments. According to this study's results, student performance improved over three assignments in every metric produced by the program. These results suggest that students using CPR became more competent at both writing and reviewing, a finding that supports previous research (Furman and Robinson, 2003; Gerdeman, Russell, and Worden, 2007; Margerum, et al., 2007; McCarty et al., 2005; Russell, 2001).

In addition to its benefit to students, CPR also provides a number of advantages to the instructor. As mentioned above, CPR provides a wealth of statistical data about student performance for each assignment. Additionally, CPR saves each student's answers to the rubric questions for every written piece they evaluate. Though not as readily accessible as the statistical data, an analysis of the rubrics can help illuminate just where students are struggling so that the instruction can be precisely targeted. Another advantage is that, although crafting new assignments in CPR requires considerable time and effort, CPR reduces the time required for grading, thus letting the instructor spend more time working closely with students and freeing TAs for other responsibilities such as facilitating laboratories or recitations. This advantage of the software is particularly relevant to large classes (Margerum et al., 2007).

Acknowledgements

We would like to thank the Texas A&M University Writing Center and the Center for Teaching Excellence for their support. We would also like to thank Dr. Victor Willson and Candace Schaefer, Associate Director of the University Writing Center. This material is based upon work supported by the National Science Foundation under Grant No. DUE-0243209.

Appendix 1. The CPR™ Process.

Step Number	Process
1	Students read the prompt provided by the instructor, access suggested resources, and submit an abstract.
2	Students use an instructor-created rubric to evaluate three abstracts created by the instructor ("calibration essays") and receive feedback on their reviews. CPR compares the students' evaluation to the instructor's evaluation of the calibration essays.
3	Students review three of their classmates' essays using the rubric introduced in step 2 and rate the essays on a scale of 1 to 10. Each student's essay is reviewed by three peers and assigned a score which is a weighted average of the three reviews.
4	Students assess their own essays using the rubric.

Appendix 2. Grading Rubric for Graders.

1. Vocabulary, Spelling and Abbreviations

Exceeds expectations	The paper contains no spelling errors. Vocabulary throughout is used properly and is appropriate to a technical audience. All non-standard abbreviations are defined.	2
Meets expectations	The paper contains no spelling errors. Vocabulary, while not used incorrectly, is not used precisely or professionally. Alternatively, paper may neglect to use technical terms when appropriate. The paper may contain one undefined, non-standard abbreviation.	1
Does not meet expectations	The paper contains spelling errors and/or mistakes in vocabulary. The paper may contain more than one non-standard abbreviation.	0

2. Grammar, Pronouns and Contractions

Exceeds expectations	The paper is free from grammatical errors. No first person plural or second person pronouns are used. The paper contains no contractions.	2
Meets expectations	The paper may contain one or two typos, but is otherwise free from grammatical errors. No first person plural or second person pronouns are used. The paper contains no contractions.	1
Does not meet expectations	The paper contains glaring grammatical errors and/or more than two typos. The paper may use inappropriate pronouns and/or contractions.	0

3. Title

Exceeds expectations	The title accurately and succinctly summarizes the exercise described in the abstract.	8
Meets expectations	The title accurately describes the exercise, but it is not succinct.	6
Does not meet expectations	The title does not accurately describe the exercise.	2

4. Background and Objectives

Exceeds expectations	The background gives accurate, concise and relevant information that places the exercise in context. The objectives for the exercise are clearly, concisely and accurately stated.	8
Meets expectations	The background gives accurate and relevant information that places the exercise in context, but may not be concise. Similarly, the objectives are clear and accurate, but not necessarily concise.	6
Does not meet expectations	The background will fail to meet expectations if it gives inaccurate and/or irrelevant information or if it fails to place the exercise in context. Objectives will fail to meet expectations if they are not accurate or clearly stated.	2

5. Methods

Exceeds expectations	The methods used in the exercise are accurately and concisely described with a level of detail appropriate to a technical audience. Moreover, only those methods that directly lead to the results reported are described.	10
Meets expectations	The methods used in the exercise are accurately described with a level of detail appropriate to a technical audience. However, the descriptions are not concise. Extraneous methods may be described that do not lead to the reported results.	7
Does not meet expectations	The methods used are not accurately described and/or the level of detail is inappropriate to a technical audience.	3

6. Results

Exceeds expectations	The important results of the exercise that lead logically to the conclusions are clearly and concisely reported using appropriate units and significant figures where appropriate.	10
Meets expectations	The important results of the exercise that lead logically to the conclusions are clearly reported. However, they may not be concise or they may use inappropriate units or significant figures.	7
Does not meet expectations	Results are reported. However, the paper may include intermediate results that do not lead directly to the conclusions and/or the results are not clearly stated.	3

7. Conclusions

Exceeds expectations	The abstract draws valid conclusions justified by the reported results in a way that is consistent with the stated objectives.	10
Meets expectations	The abstract draws valid conclusions justified by the reported results. However, the conclusions do not necessarily parallel the objectives. Alternatively, the paper may neglect conclusions suggested by the results	7
Does not meet expectations	The abstract contains conclusions that are not justified by the results.	3

Appendix 3. Explanation of CPR-generated scores.

Score	Explanation
Text rating (TR)	Text rating is a weighted average of scores given by three peer reviewers. Weighting is based on reviewing competency of the peer (see RCI). Peer reviewers are instructed to base the score on analysis guided by the calibration questions. Since the calibration questions include both content-related questions and writing-related questions, TR can reflect both content understanding and writing competence.
Calibration score	The student's calibration score is computed by comparing (for each of the three sample essays) the student's responses to the calibration questions to the instructor's responses and the student's text rating to the instructor's. The instructor determines what % of the style questions, % of the content questions must be correct, and what deviation from the instructors' text rating is allowable, in order to receive credit for review of each calibration essay.
Calibration Deviation	Calibration deviation refers to the difference between the student's rating of each sample essay with the instructor's rating.
Percent correct style and Percent correct content	For each set of calibration questions, the instructor labels some as style questions and some as content questions. For each sample essay, CPR compares student answers to the calibration questions with instructor answers and determines % correct in the style category and % correct in the content category.
Reviewer competency index (RCI)	The reviewer competency index is computed (by the CPR program) following student review of three instructor-provided essays. RCI computation uses a comparison of student and instructor responses to calibration questions as well as of student and instructor global rating of the essays.
Review score	The student's review score is based on a comparison of the student's rating of the peer's text with the weighted average of all three student reviewers' ratings. The instructor determines how small the deviation from the weighted average must be in order for the student to receive full or partial credit for the review phase.
Review Deviation	Review deviation refers to the difference between the student's rating of a peer's text with the weighted average of the ratings given by all three students to whom that text was assigned.
Self-Assessment score (SA)	CPR computes each student's self-assessment score by comparing the global rating student gives his/her own text to the weighted average of the text ratings assigned by peers (see TR). The instructor determines how small the deviation from the weighted average must be in order for the student to receive full or partial credit for the self-assessment phase.
Overall grade	The student's overall grade for a CPR assignment is computed from four elements: (1) text rating (2) calibrations (3) reviews (4) self-assessment. The instructor determines the weight given to each of the four elements.

References

Barnett, R**.** W., and Blumner, J. S. (Eds). (1999). *Writing centers and writing across the curriculum programs.* Westport, CT: Greenwood Press.

Boud, D. (1990). Assessment and the promotion of academic values. *Studies in Higher Education, 15*(1), 101-111.

Brindley, C., and Scoffield, S. (1998). Peer assessment in undergraduate programmes. *Teaching in Higher Education, 3*(1), 79–89.

Bunderson, E., and Christensen, M. E. (1995). An analysis of retention problems for female students in university computer science programs. *Journal of Research on Computing in Education, 28*(1), 1-18.

Camp, T. (1997). The incredible shrinking pipeline. *Communication of the ACM, 40,* 103.

Cutler, H., and Price, J. (1995). The development of skills through peer assessment. In A. Edwards and P. Knight (Eds.), *Assessing Competence in Higher Education* (pp. 150-159). London: Kogan Page.

Davies, V., and Camp, T. (2000). Where have women gone and will they be returning? *CPSR Newsletter, 18*(1).

Dochy, F., Segers, M., and Sluijman, S. (1999). The use of self-, peer and co-assessment in higher education: A review. *Studies in Higher Education, 24,* 331-350.

Doornekamp, B. G. (1993). Students' valuation of the use of computers in education. *Computers and Education, 21*(1/2), 102-113.

Durndell, A., and Thomson, K. (1997). Gender and computing: A decade of change? *Computers and Education, 28*(1), 1-9.

Falchikov, N. (1995). Peer feedback marking: Developing peer assessment. *Innovations in Education and Training International, 32*(2), 175-187.

Freeman, M. (1995). Peer assessment by groups of group work. *Assessment and Evaluation in Higher Education, 20*(3), 289-301.

Furman, B., and Robinson, W. (2003). *Improving engineering report writing with Calibrated Peer Review*. Paper presented at the 33rd ASEE/IEEE Frontiers in Education Conference, November 5-8, 2003, Boulder, CO, pp. F3E-14-F3E-15.

Gerdeman, R. D., Russell, A. R., and Worden, K. J. (2007). Web-based student writing and reviewing in a large biology lecture course. *Journal of College Science Teaching* (March/ April 2007), 46-52.

Herrington, A. J. (1997). Developing and responding to major writing projects. In Sorcinelli, M. D., and Elbow, P. *New Directions for Teaching and Learning: Writing to learn: strategies for assigning and responding to writing across the disciplines (69).* San Francisco: Jossey-Bass.

Holliday, W.G., Yore, L. D., and Alvermann, D.E. (1994). The reading-science learning-writing connection: Breakthroughs, barriers, and promises. *Journal of Research in Science Teaching, 31*, 877-894.

Jannsen Reinen, I., and Plomp, T. (1997). Information technology and gender equality: A contradiction in terminis. *Computers in Education, 28*(2), 65-78.

Klein, P. D. (1999). Reopening inquiry into cognitive processes in writing-to-learn. *Educational Psychology Review, 11*(3), 203-270.

Kovac, J., and Sherwood, D. W. (1999). Writing in chemistry: An effective learning tool. *Journal of Chemical Education, 76*(10), 1399-1403.

Lloyd-Jones, R. (1977). Primary-trait scoring. In C. Cooper and L. Odell (Eds). *Evaluating Writing: Describing, measuring, judging* (pp. 33-66). Urbana, IL: National Council of Teachers of English.

Lowman, J. (1996). Assignments that promote learning. In Menges, R. J., Weimer, M., and Associates, *Teaching on solid ground: Using scholarship to improve practice.* San Francisco: Jossey-Bass.

Margerum, L. D., Gulsrud, M., Manlapez, R., Rebong, R., and Love, A. (2007). Application of calibrated peer review (CPR) writing assignments to enhance experiments with an environmental chemistry focus. *Journal of Chemical Education, 84*(2), 292-295.

McCarty, T., Parkes, M. V., Anderson, T. T., Mines, J., Skipper, B. L., and Greboksy. (2005). Improved patient notes from medical students during web-based teaching using faculty-calibrated peer review and self-assessment. *Acad Med*, *80*, 67-70.

Odell, L. (1992). Context-specific ways of knowing and the evaluaton of writing. In A. Herrington and C. Moran (Eds). *Writing, teaching and learning in the disciplines* (pp. 86-98). NY: Modern Language Association.

Orsmond, P., Merry, S., and Callaghan, A. (2004). Implementation of a formative assessment model incorporating peer and self-assessment. *Innovations in Education and Teaching International, 41*(3), 273-290.

Paul, R. (1995). *Critical thinking: How to prepare students for a rapidly changing world.* Santa Rosa, CA: The Foundation for Critical Thinking.

Pope, N. K. (2005). The impact of stress in self- and peer assessment. *Assessment and Evaluation in Higher Education, 30*(1), 51-63.

Reese-Durham, N. (2005). Peer evaluation as an active learning technique. *Journal of Instructional Psychology, 32*(4), 338-345.

Rivard, L. P., Stanley, B., and Straw, S. B. (2000). The effect of talk and writing on learning science: An exploratory study. *Science Education, 84*(5), 566-593.

Russell, A. (2001). *The evaluation of CPR. Prepared for HP e-Education; Business Development.* Los Angeles: UCLA.

Saavedra, R., and Kwun, S. K. (1993). Peer evaluation in self-managing work groups. *Journal of Applied Pyschology, 78*(3), 450-462.

Searby, M., and Ewers, T. (1997). An evaluation of the use of peer assessment in higher education: A case study in the school of music, Kingston University. *Assessment and Evaluation in Higher Education, 22*(4).

Sluijsmans, D., Brand-Gruwel, S., Van Merrienboer, J. (2002). Peer assessment training in teacher education. *Assessment and Evaluation in Higher Education, 27*(5), 443-454.

Sluijsmans, D., Dochy, F., and Moerkerke, G. (1999). Creating a learning environment by using self-, peer- and co-assessment. *Learning Environments Research, 1*, 293-319.

Sobral, D. T. (1997). Improving learning skills: A self-help group approach. *Higher Education, 33*, 39-50.

Stefani, A. J. (1992). Comparison of collaborative, self, peer, and tutor assessment in a biochemistry practical. *Biochemical Education, 20*, 148-151.

Stefani, L. A. J. (1994). Peer, self and tutor assessment: Relative reliabilities. *Studies in Higher Education, 19*(1), 69-75.

Sternberg, R. J. (1994). Diversifying instruction and assessment. *The Educational Forum, 59*(1), 47-52.

Sutton, R. (1991). Equity and computers in the schools: A decade of research. *Review of Educational Research, 61*(4), 475-503.

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

Volman, M. (1997). Gender-related effects of information and computer literacy education. *Journal of Curriculum Studies, 29*(3), 315-328.

Wen, M. L., and Tsai, C. (2006). University students' perceptions of and attitudes toward (online) peer assessment. *Higher Education, 51*(1), 27-44.

Wheeler, E. D., Balazs, G. G., and McDonald, R. L. (1997). *Writing as a teaching and learning tool in engineering courses.* Proceedings of the 1997 ASEE/IEEE Frontiers in Education Conference, IEEE catalog number 97CH36099, pp. 1538-1542.

Williams, E. (1992). Student attitudes towards approaches to learning and assessment. *Assessment and Evaluation in Higher Education, 17*, 45–58.

Wright, W. A., Herteis, E. M., and Abernethy, B. (2001). *Learning through writing: a compendium of assignments and techniques* (revised). Halifax, Canada: Dalhousie University.

Young, B. J. (2000). Gender differences in student attitudes toward computers. *Journal of Research on Computing in Education, 33*(2), 204-216.