Journal of the Scholarship of Teaching and Learning, Vol. 8, No. 2, May 2008, pp. 25-37.

Effectiveness of Calibrated Peer Review™ for improving writing and critical thinking skills in biology undergraduate students

Adalet Baris Gunersel, Nancy J. Simpson, Karl J. Aufderheide, and Li Wang[1]

Abstract: This study focuses on student development with Calibrated Peer Review (CPR)™, a web-based tool created to promote writing and critical thinking skills. Research questions focus on whether or not students showed improvement in writing and reviewing competency with repeated use of CPR in a senior-level biology course and whether the difference between higher performing and lower performing students decreased over time. Four repeated measures analyses were conducted with different sets of students. Repeated measures analyses indicate that students showed improvement in writing skills and reviewer competency with repeated use of CPR. The difference between higher and lower performing students decreased over time in both writing skills and reviewer competency.

Keywords: science education, critical thinking, innovative teaching tools, writing skills, peer review, undergraduate education.

Calibrated Peer Review (CPR)™ is a web-based tool for authoring and managing student writing assignments (for more information, see http://cpr.molsci.ucla.edu/). CPR assignments engage students in writing and in reviewing their peers' work, and include a calibration phase during which students practice reviewing according to an instructor-designed rubric. While there are a few published studies that provide evidence of the value of CPR as a tool for improving students' conceptual learning, as well as their writing and critical thinking skills (e.g., Furman and Robinson, 2003; Gerdeman, Russell, and Worden, 2007; Margerum, Gulsrud, Manlapez, Rebong, and Love, 2007; McCarty et al., 2005) more research is needed. What are the characteristics of effective CPR assignments? Is CPR effective for all students? What strategies for implementation lead to success? Questions such as these intrigued a biology instructor and two faculty developers, all of whom had been working on an NSF-funded project focusing on CPR. This joint curiosity led to the current study investigating student outcomes in three semesters (Spring 2005, Spring 2006, and Spring 2007) of a senior-level biology course by using repeated measures analyses of CPR-generated data. The focus is the effectiveness of repeated use of CPR for improving student writing and reviewing competency in biology with CPR. The course's instructor, who was a part of the research team, provides information regarding the course to give a context for the study. This study adds to the literature about the effectiveness of CPR by investigating the development of student writing in biology and reviewing competency in a senior-level biology class. Specifically, there are three research questions:

1. Did repeated use of CPR in a senior-level biology course result in improvement in writing and reviewing skills of initially lower performing students?

[1] Center for Teaching Excellence, 533 Blocker, MS 4246, Texas A&M University, College Station, TX 77843, bgunersel@tamu.edu; n-simpson@tamu.edu; kauf@mail.bio.tamu.edu; wangli_99@yahoo.com

2. Did the difference between higher performing and lower performing students decrease with repeated use of CPR?
3. Did repeated use of CPR in a senior-level biology course result in improvement in writing and reviewing skills in general?

I. Calibrated Peer Review (CPR)™.

In order to provide a greater understanding of the study, further explanation of CPR and research related to CPR are presented in this section. Developed at UCLA for the Molecular Science Project, one of the NSF-supported Chemistry Systemic Reform Initiatives, CPR was designed to give students practice in writing and peer review, since both are expected competencies in scientific fields (Russell, 2001).

One of CPR's aims is to develop students' skills of discipline-specific writing, which has become prominent in education (Emerson, MacKay, MacKay, and Funnell, 2006; Lea and Street, 1998). The underlying the pedagogy of CPR is reinforced by numerous studies supporting the educational value of both writing (Holliday, Yore, and Alvermann, 1994; Klein, 1999; Kovac and Sherwood, 1999; Lowman, 1996; Rivard, Stanley, and Straw, 2000) and peer review (Falchikov, 1995; Orsmond, Perry, and Callaghan, 2004; Searby and Ewers, 1997; Sluijsmans, Brand-Gruwel, and Van Merrienboer, 2002; Sluijmans, Docky, and Moerkerke, 1999; Topping, 1998). Although peer review may be a source for hesitation among students, several studies suggest that peer review can be as reliable as faculty assessment (Falchikov, 1995; Freeman, 1995; Saavedra and Kwun, 1993; Sluijsmans et al., 1999; Stefani, 1994; Topping, 1998).

In addition to having students write and review peers' work, CPR has the students practice reviewing in the "calibration phase." In order to create a CPR assignment, instructors produce the following components:

Instructions for writing. Instructions include suggested resources, questions to guide student thinking, and a "writing prompt" that tells students such things as the topic, format and audience for their writing.

Calibration questions. A set of questions that direct students attention to content and style characteristics of a completed assignment and form the basis for assigning a text rating.

Three sample essays. High, average, and low quality essays that are the responses to the assignment. (Instructors review and rate these essays using the calibration questions.)

Student work on a CPR assignment occurs in three phases:

Text entry phase. Students read instructions, access suggested resources, and write and submit their essays.

Calibration phase. Students are presented with the three sample essays, along with the calibration questions. For each essay, students answer the calibration questions and assign a rating. CPR assigns a reviewer competency index based on a comparison of the student review to the instructor review of each essay;

Review phase. Students are presented first with three classmates' essays (randomly assigned and anonymous) and then with their own essay, all of which they review and rate using the same set of calibration questions.

Instructor-reported experiences and a limited number of studies have suggested that CPR 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 suggested that students felt they were becoming "better technical reviewers" with CPR assignments (p. 294). They also found that students mastered the class content through both the calibration phase and the review phase. Pelaez (2002) compared the learning outcomes of undergraduate nonscience majors taught with lectures and taught with CPR™ in an introductory physiology course. The results suggested that the performance of students who had completed problem-based learning assignments in CPR was better than or equal to the performance of students who had received "traditional instruction" (p. 181). Pelaez (2002) noted:

> The favorable results may be a product of the work students complete when writing about their thinking, or perhaps students did better because PW-PR (problem-based writing with peer review) made it possible for them to confront and resolve difficulties they encountered relating concepts. (p. 181)

II. The Context of the Study.

Data from students in three semesters (Spring 2005, Spring 2006, and Spring 2007) of a senior-level biology course were used. Each semester, students completed the same four CPR assignments and three highest scores counted for the final grade. The assignments were ordered in increasing difficulty: "Why Do We Use The SI System Of Measurement In Science?" "Mitosis Through the Microscope: Advances in Seeing Inside Live Dividing Cells," "Microtubules and Motor Proteins," and "Cajal Bodies and Coilin—Moving Towards Function." While the first one was an example assignment from the CPR assignment library, the other three were created by the instructor.

This ordering made the assignments get "more focused on a specific area of cell biology and much more detailed in the kinds of information a student would have to collect and condense into a series of paragraphs." The instructor used the assignments for a dual purpose: They were related to lecture topics and there was "a sequence of increasing complexity and specific focus as to the nature of the information that they're going to have to deal with."

The second assignment (in 2005 and 2006) ("Mitosis Through the Microscope: Advances in Seeing Inside Live Dividing Cells") was a historical overview of how a specific microscope has been used in cell biology. The article that the students had to work with to complete the assignment was a "general article, an overview" that was "roughly coordinated to some of the classes they did in the beginning of the semester." The third assignment ("Microtubules and Motor Proteins") was "much more detailed about a specific set of cellular structure and motor proteins that interact with them." The structure of this assignment was slightly different: Students didn't have one article as a source, rather they were linked to a series of research websites. This assignment was more difficult than the second, since students were "doing much more of a diffuse search to several sources of information." The fourth assignment ("Cajal Bodies and Coilin—Moving Towards Function") was the most difficult of all, partly because of the topic of the assignment, but also but also because the source material (the review article) they read was not very well written and was poorly organized. Therefore to answer the guiding questions, they had to read the entire essay and select information from paragraph to paragraph to construct a comprehensive narrative. So, in fact, students had to write an essay that was better organized than the original source material.

III. Methods.

Specifically, the study addresses three questions:

1. Did repeated use of CPR in a senior-level biology course result in improvement in writing and reviewing skills of initially lower performing students?
2. Did the difference between higher performing and lower performing students decrease with repeated use of CPR?
3. Did repeated use of CPR in a senior-level biology course result in improvement in writing and reviewing skills in general?

Data from students in three semesters (Spring 2005, Spring 2006, and Spring 2007) of a senior-level biology course were used. For the analyses two CPR-generated scores, reviewer competency index (RCI) and text rating (TR), were included as dependent variables. The reviewer competency index (RCI) 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 instructor-generated calibration questions, as well as of student and instructor global rating of the essays. Text rating (TR), on the other hand, is a weighted average of scores given by three peer reviewers. Weighting is based on reviewing competency (RCI) of the peer. 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. In summary, TR is used as a measure of writing quality and content understanding, while RCI is used as a measure of students' ability to review. For each CPR assignment students receive a TR ranging from 1 to 10 and a RCI ranging from 1 to 6. Students who had completed fewer than three of the four assignments were eliminated from the analysis.

Students were categorized into two groups according to their TR and RCI scores from the first assignment: higher performing (third quartile; highest 25%) and lower performing (first quartile; lowest 25%) (see Table 1). The second quartile was eliminated in order to focus on the development of higher and lower performing students. Thus, TR scores of 47 students (18 from Spring 2005, 15 from Spring 2006, and 14 from Spring 2007) were included and RCI scores of 83 students (27 from Spring 2005, 26 from Spring 2006, and 30 from Spring 2007) were included (see Table 1). This discrepancy between the numbers occurred since the second quartile was larger for TR scores than RCI scores.

In addition to this, all students regardless of performance level were included in separate repeated measure analyses. Table 4 presents the number of students for each assignment.

Table 1. Higher and lower performing groups, 2005-2006-2007.

	Lower Performing	Number	Higher Performing	Number	All students
TR1	≤6.4925	23	≥8.9600	24	47
RCI1	≤2.000	45	≥6.000	38	83

IV. Data Analysis.

Four repeated measures analyses were conducted, two of which focused on students at initial performance levels and two of which included all students, regardless of performance level. The first analysis included the TR scores of a total of 47 students. The second analysis

included the RCI scores of a total of 83 students (27 from Spring 2005, 26 from Spring 2006, and 30 from Spring 2007) from groups of higher and lower performance. Both analyses included lower performance and higher performance as the grouping variable and the number of assignments (4) as the within-subjects factor.

The other two repeated measures analyses included all students regardless of performance level. Students' TR and RCI scores were used as dependent variables, the semester as the grouping variable, and the number of assignments (4) as the within-subjects factor.

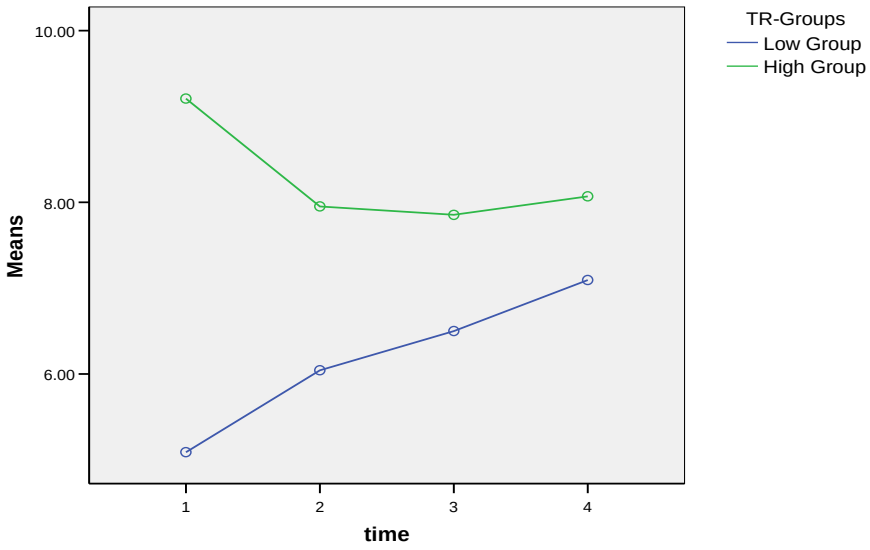
V. Results.

When considering initially higher and lower performing students in TR, although there was no overall statistically significant change over four assignments (df= 3, F= 1.813, p< 0.149), the change of means of higher and lower performing students was statistically significant over four assignments at alpha level .01 (df= 3, F=14.370, p< 0.000). The mean for the lower performing group increased steadily throughout the semester, while the higher performing group's mean decreased (see Table 2 and Graph 1). Also, the difference between the groups decreased throughout the semester.

Table 2. TR Descriptive Statistics for Spring 2005-2006-2007.

		M	*SD*
TR1	Higher performing	9.2079	.30570
	Lower performing	5.0876	1.04912
	All students	7.0448	2.22479
TR2	Higher performing	7.9521	1.45644
	Lower performing	6.0419	1.65149
	All students	6.9493	1.81961
TR3	Higher performing	7.8542	1.03237
	Lower performing	6.4995	1.56828
	All students	7.1430	1.49083
TR4	Higher performing	8.0689	.96489
	Lower performing	7.0938	1.02561
	All students	7.5570	1.10106

Graph 1. Spring 2005-2006-2007 means plot of TR Groups.

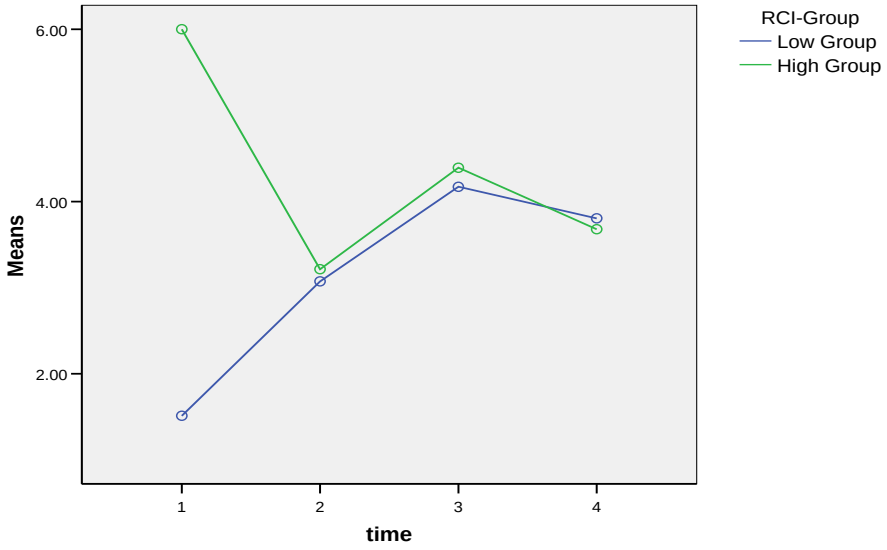


When considering initially higher and lower performing students in RCI, there was overall statistically significant change four assignments at alpha level 0.01 (df= 3, F= 8.479, p< 0.000) and significant change in higher and lower performing students at alpha level 0.01 (df= 3, F= 47.829, p< 0.000). The lower performing students' group improved throughout the semester, while the higher performing group fluctuated (see Table 3 and Graph 2). However, the means of both groups decreased in the fourth, most difficult, assignment. All students together showed a statistically significant increase from the first assignment to the last (means of 3.3333 to 3.7536) and the difference between the groups had almost disappeared by the second assignment (means of 3.6786 and 3.8049).

Table 3. RCI Descriptive Statistics for Spring 2005-2006-2007.

		M	*SD*
RCI1	Higher performing	6.000	0.0000
	Lower performing	1.5122	0.77852
	All students	3.3333	2.29876
RCI2	Higher performing	3.2143	1.54817
	Lower performing	3.0732	1.43858
	All students	3.1304	1.47442
RCI3	Higher performing	4.3929	1.59488
	Lower performing	4.1707	1.56369
	All students	4.2609	1.56855
RCI4	Higher performing	3.6786	1.46701
	Lower performing	3.8049	1.70616
	All students	3.7536	1.60336

Graph 2. Spring 2005-2006-2007 means plot of RCI Groups.



When the TRs of all students were included, there was a statistically significant change over four assignments at alpha level 0.05 (df= 3, F= 2.814, p< 0.041). However, there was no statistically significant change when separated according to different semesters (df= 6, ,F= 0.888, p< 0.0506). While the mean of the TR scores initially decreased from the first assignment to the second, they increased steadily from the second to the fourth assignment (see Table 4).

Table 4. TR Descriptive Statistics (semesters combined).

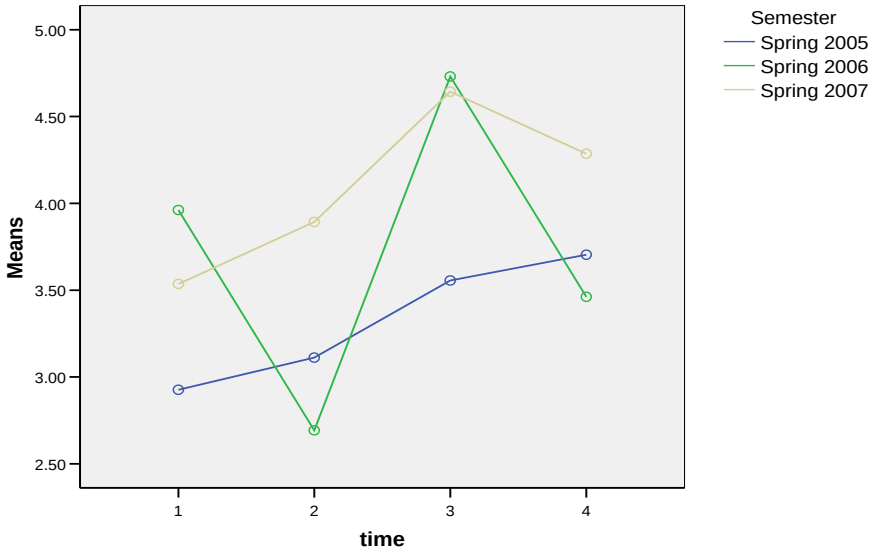
Assignment	*# of Students*	*Minimum*	*Maximum*	*M*	*SD*
1	94	2.89	10.00	7.5089	1.62404
2	93	3.12	11.98	7.2424	1.51903
3	87	3.44	10.00	7.3366	1.39329
4	78	4.06	9.52	7.6663	1.09819

When the RCIs of all students were included, there was statistically significant change over four assignments at alpha level 0.01 (df= 3, F= 6.709, η^2=0.088, p< 0.000). There was also statistically significant change when separated according to different semesters at alpha level 0.05 (df= 6, F= 5.871, p< 0.042). While RCIs slightly declined from the first assignment to the second, they increased from the second to the third, but then decreased again on the fourth, most difficult assignment (Table 5). While students in Spring 2005 showed constant improvement in their RCIs, students' means in Spring 2006 fluctuated (see Table 5 and Graph 3). Students in Spring 2007 showed improvement during the first three assignments, and then a decline. However, the mean on the final assignment was higher than the mean on the first one.

Table 5. RCI Descriptive Statistics.

		M	*SD*
RCI1	Spring 2005	2.9259	2.12903
	Spring 2006	3.9615	2.48967
	Spring 2007	3.5357	1.75293
	Combined	3.7238	2.06386
RCI2	Spring 2005	3.1111	1.18754
	Spring 2006	2.6923	1.43581
	Spring 2007	3.8929	1.87260
	Combined	3.6058	1.75940
RCI3	Spring 2005	3.5556	1.57708
	Spring 2006	4.7308	1.56353
	Spring 2007	4.6429	1.39348
	Combined	4.4796	1.58751
RCI4	Spring 2005	3.7037	1.83586
	Spring 2006	3.4615	1.44861
	Spring 2007	4.2857	1.54750
	Combined	3.8876	1.51836

Graph 3. Means plot for RCI.



VI. Conclusions.

Our study suggests that repeated practice with CPR is an effective way to help students develop writing and reviewing skills in biology, supporting other studies that have found CPR's usefulness with writing and reviewing skills (e.g., Furman and Robinson, 2003; Gerdeman, Russell, and Worden, 2007; Margerum et al., 2007; McCarty et al., 2005; Pelaez,

2002). Since the instructor-created rubric included both content-related criteria and writing-related criteria, improved TRs indicate increased student ability both to understand the content focus of the CPR assignments and to write about this content in a coherent manner. In addition to this, results showed that the difference between the higher performing and lower performing students decreased in both TRs and RCIs.

Repeated use of CPR appears to be particularly beneficial for initially lower performing students. In our study, the students who did poorly on the first assignment exhibited statistically significant improvement with repeated use of CPR in both TR and RCI. While the difficulty of the final assignment did not impact TRs, it did impact RCIs, as student scores showed a decrease. Still, the RCIs of the lower performing students on the final assignment were higher than those on the first assignment. This improvement occurred despite the fact that grading rubric was different for each assignment. The students appeared to have become more adept at, as the instructor put it, "internalizing" a set of criteria for evaluation.

Initially higher performing students showed a slight, but significant, decrease in their TRs (Graph 1) which could have two reasons: Since three of the highest scores counted, students' efforts may have decreased, or it may have been the "regression to the mean" which suggests that students' initially high scores would be more likely to decrease. These students fluctuated in their RCIs (Graph 2). Just like the lower performing students, their RCIs were impacted by the increased difficulty of the fourth assignment, while their TRs were not. A possible reason for this finding may be that the difficulty of the assignment impacted students' ability to match the evaluations of the instructor, which is what the RCI is based on. Another possible reason is the following: The difficulty of the fourth assignment rose from its text which was not well written. Thus, it may have been more difficult for the students to examine the details of the text and rate their peers in a way that is similar to the instructor's rating, which impacts their RCIs. On the other hand, the difficulty of the text may not have impacted students' ability to get enough information to write a medium- to high-quality essay, which leads to TRs remaining unaffected by the difficulty.

The results for the initially higher and lower performing student groups are consistent with the findings of Gerdeman, Russell, and Worden (2007): Students with the lowest initial levels of performance gained the most over time, while students with the highest levels of performance slightly declined.

Analysis including all students regardless of performance level also bore interesting results. There was a significant change in TRs over four assignments with a steady increase after the second assignment. There was no significance when students were separated according to semesters possibly due to the small sample size in each semester (22 from Spring 2005, 20 from Spring 2006, and 28 from Spring 2007).

Changes in RCIs in each semester varied and were statistically significant. While in Spring 2005, students' RCIs showed a statistically significant increase, in Spring 2006 scores fluctuated, with means decreasing from the first to the second assignment, then increasing from the second to the third, and decreasing again from the third to the fourth. Students' scores in 2007 showed a statistically significant increase over three assignments, and then a decline. This decline is not unexpected as the fourth assignment was the most difficult. In 2005 and 2007, RCIs on the final assignment were higher than the initial, which was not the case in 2006.

In addition to this, CPR brought students with initially different levels of performance closer together in their scores: for both variables, TR and RCI, the difference between lower performing and higher performing students decreased over four assignments, which can be

observed in Graph 1 and 2. The initial differences in performance could be attributed to levels of preparation and ability which the students brought to class. CPR seems to be a useful tool that helps students overcome initial shortcomings and brings students together in skills.

It should also be noted that both the advancement of students who were at a lower performing level in the first assignment and the decline of the differences between students at different levels of abilities took place without a significant amount of feedback from the instructor outside of the CPR program. This is an important aspect of CPR as it which is frees time for other instructional tasks and gives instructors with large classes the opportunity to use writing.

In today's learning environment where it is important to be able to critique and to probe, CPR shows promise as a learning tool that gives students the opportunity to exercise their writing and critical thinking skills and opens new avenues to learning. Our interview with the instructor revealed that these aspects of CPR were his reasons to continue using this educational tool, although it was unfamiliar for both him and the students. He believed that college students needed further experience in writing and reviewing—using specific grading standards—which they would need in the future.

This study was a retrospective analysis and not an a priori designed experiment. It addressed questions regarding the effect of CPR and student learning, and used instructor reflection to interpret data generated by CPR. Since CPR assignments are discipline-specific, and implementation of CPR assignments is strongly influenced by the context and structure of the course, the accumulation of studies in a range of disciplines and contexts will be needed for greater understanding of factors influencing the effectiveness of CPR as an educational tool. A future study may also include interviews with the students in order to get an understanding of their experience and perspective.

Acknowledgements

We wish to acknowledge the consultation and feedback from Dr. Victor Wilson and Dr. Stephanie Knight, Department of Educational Psychology, Texas A&M University, and Dr. Arlene Russell, Department of Chemistry, University of California, Los Angeles in preparing this manuscript. This material is based upon work supported by the National Science Foundation under Grant No. DUE-0243209.

References

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

Cross, K. P. **(**1998). Classroom research: Implementing the scholarship of teaching. In Angelo, T. (Fall, 1998). *New Directions for Teaching and Learning: Classroom assessment and Research: an update on uses, approaches, and research findings (75).* San Francisco: Jossey-Bass.

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.

Educause Learning Initiative (September 2005). Calibrated Peer Review: A writing and Critical-thinking instructional tool. *Innovations and Implementations: Exemplary Practices in Teaching and Learning.* UCLA, USC. Retrieved October 15th, 2005 from http://www.educause.edu/ir/library/pdf/ELI5002.pdf.

Ellis, G. (2001). Looking at ourselves – self-assessment and peer assessment: Practice examples from New Zealand. *Reflective Practice, 2*(3), 289-302.

Emerson, L., MacKay, B. R., MacKay, M. B., and Funnell, K. A. (2006). A team of equals: Teaching writing in the sciences. *Educational Action Research, 14*(1), 65-81.

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.

Goody, J. (1994). *Entre l'oralite´ etl'e´criture.* Paris: Presses universitaires de France.

Halliday, M. A. K., and Martin, J. R. (1993). *Writing science: Literacy and discursive power.* Pittsburgh, PA: University of Pittsburgh Press.

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.

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.

Langer, J. A., and Applebee, A. N. (1987). *How writing shapes thinking* (Research Report No.

22). Urbana, IL: National Council of Teachers of English.

Lea, M. R., and Street, B. V. (1998). Student writing in higher education: An academic literacies approach. *Studies in Higher Education, 23*(2), 157-172.

Liu, J., Pysarchik, D. T. and Taylor, W. (2002). Peer review in the classroom. *BioScience, 52*(9), 824-829.

Lowman, J. (1996). Assignments that promote learning. In R. J. Menges, M. Weimer, and Associates (Eds.), *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.

McGinley, G. A., and Tierney, R. J. (1989). Traversing the topical landscape: Reading and writing as ways of knowing. *Written Communication, 6*, 243–269.

McKeachie, W. (2002). *McKeachie's teaching tips: Strategies, research, and theory for college and university teachers* (11th ed.). Boston: Houghton Mifflin Co.

National Council of Teachers of Mathematics. (1993). *Assessment Standards for School Mathematics: Working Draft*, Reston, VA: NCTM.

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.

Pelaez, N. J. (2002). Problem-based writing with peer review improves academic performance in physiology. *Advanced Physiology Education, 26*, 174-184.

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

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 Psychology, 78(3), 450-462.

Shafer, J.L. (1997). Software for multiple imputation. University Park, PA: The Pennsylvania State University Department of Statistics.

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

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

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

Topping, K. J., Smith, E. F., Swanson, I., and Elliot, A. (2000). Formative peer assessment of academic writing between postgraduate students. *Assessment and Evaluation in Higher Education, 25*(2), 149-169.