

ARTICLE

Writing to Learn: An Evaluation of the Calibrated Peer Review™ Program in Two Neuroscience Courses

J. Roxanne Prichard

Psychology Department and Neuroscience Program, Bates College, Lewiston, ME 04240

Although the majority of scientific information is communicated in written form, and peer review is the primary process by which it is validated, undergraduate students may receive little direct training in science writing or peer review. Here, I describe the use of Calibrated Peer Review™ (CPR), a free, web-based writing and peer review program designed to alleviate instructor workload, in two undergraduate neuroscience courses: an upper-level sensation and perception course (41 students, three assignments) and an introductory neuroscience course (50 students; two assignments). Using CPR online, students reviewed primary research articles on assigned 'hot' topics, wrote short essays in response to specific guiding questions, reviewed standard 'calibration' essays, and provided anonymous quantitative and qualitative peer reviews. An automated grading system calculated the final scores based on a student's essay quality (as determined by the average of three peer reviews) and his or her accuracy in evaluating 1) three standard calibration

essays, 2) three anonymous peer reviews, and 3) his or her self review. Thus, students were assessed not only on their skill at constructing logical, evidence-based arguments, but also on their ability to accurately evaluate their peers' writing. According to both student self-reports and instructor observation, students' writing and peer review skills improved over the course of the semester. Student evaluation of the CPR program was mixed; while some students felt like the peer review process enhanced their understanding of the material and improved their writing, others felt as though the process was biased and required too much time. Despite student critiques of the program, I still recommend the CPR program as an excellent and free resource for incorporating more writing, peer review, and critical thinking into an undergraduate neuroscience curriculum.

Keywords: peer review, writing to learn; web-based learning; learning technology; Calibrated Peer Review

Historically, the majority of undergraduate science courses are taught through didactic lectures and are evaluated using multiple choice and short answer formats. Incorporating writing into the course structure deepens the understanding of major concepts by requiring that students comprehend, analyze and synthesize information beyond the level of knowledge recall (Bloom 1956; Bean, 1996). Peer review requires additional skill sets (e.g., evaluation, judging against standards, justification of judgments) that further enrich students' understanding of core concepts (Committee on Undergraduate Science Education, 1997; Rao and DiCarlo, 2000).

Although peer review is virtually the only way higher-level science information is critiqued and validated, it is not always taught in a college environment. The time and workload associated with orchestrating in-class peer reviews and evaluating both essays and peer reviews deter many educators from adopting this pedagogical tool, especially in large courses. Furthermore, peer reviews performed by untrained students can often vary in validity, reliability, and usefulness.

To address these concerns, Orville Chapman, the late Dean of Educational Innovation, Professor of Organic Chemistry, and Principal Investigator of the Molecular Science Project at UCLA, developed Calibrated Peer Review (CPR), a program designed to strengthen students' skills in writing and peer review, while empowering instructors to assign frequent writing assignments without increasing grading workload. CPR, funded by the National Science Foundation and by the Howard Hughes Medical

Institute, is an internet-based pedagogical tool that manages the electronic text entry and the anonymous review process, analyzes and assesses student input, and prepares summary reports for the instructor and students (Chapman, 1999). CPR is designed to minimize the traditional problems with peer assessment through 1) the use of clear criteria of grading parameters, 2) a standardized 'calibrating' practice session in peer review, 3) double anonymity of reviewers and reviewees, and 4) including multiple reviewers of each essay. The CPR user-base is growing exponentially; as of 2004, CPR was being used in over 1900 different courses ranging in class size from 20 to 500, at more than 500 academic institutions across the country (Russell, 2004). Here, I summarize the use of CPR in two undergraduate neuroscience courses by discussing the format of the CPR program, the demands on instructor, student performance on assignments, and student assessment of the program.

MATERIALS AND METHODS

Courses

CPR was used in two neuroscience courses: an introductory neuroscience course (50 students, 86% first-years and sophomores) and an upper level sensation and perception course (40 students, 95% juniors and seniors). In the introductory course, students wrote one summary essay based on a scholarly review and one argumentative essay based on a primary research article. In the upper level course, students composed two summary and

synthesis essays and one argumentative essay, all of which were based on primary research articles of increasing sophistication. Information on the sample groups, student performance, and instructor grading demands is summarized in Table 1.

| Sample Group | Intro Level | Upper Level |
|--|-------------|-------------|
| Students enrolled (n) | 50 | 40 |
| Completed evaluations (n) | 39 | 31 |
| 1st & 2nd yr. students (%) | 86% | 5% |
| 3rd & 4th yr. students (%) | 14% | 95% |
| Summary/Synthesis Writing Assignments (n) | 1 | 2 |
| Argumentative Writing Assignments (n) | 1 | 1 |
| Student Performance | | |
| Mean text rating (1-10 scale) | 7.55 | 7.33 |
| Mean deviation (calibration essays) | 1.04 | 1.37 |
| Mean deviation (peer reviews) | 0.88 | 0.93 |
| Mean deviation (self reviews) | 1 | 1.09 |
| Instructor Grading Demands | | |
| % of text grades adjusted per assignment | 21% | 32% |
| % of overall grades adjusted per assignment | 45% | 70% |
| Students who revised essays for extra credit | 63% | 40% |

Table 1. Demographic data and student performance on CPR assignments. CPR assignments were used in an introductory neuroscience course and in an upper level sensation and perception course. The mean text rating, generated by the weighted average of three anonymous peer reviews, was usually consistent with the instructor's evaluation of the student essays. However, the overall assignment grade, as calculated by the mean text rating, and the mean deviations on calibrations, peer reviews and self reviews, required grade adjustments by the instructor more frequently.

Before the first CPR assignment, students completed an anonymous survey about their previous experience with peer review. At the end of the semester, students completed another anonymous survey that asked for qualitative and quantitative feedback on their experiences with the CPR program. Specifically, students were asked to state what they liked most and least about the CPR program, and to rate on a 1-5 Likert scale how much they felt that their writing and peer review skills improved over the semester, how often they provided and received helpful peer reviews, to what degree they felt that they were evaluated fairly by the CPR program, and how much work they put into CPR assignments, compared to writing assignments of equal length.

Creating assignments

The CPR program offers exceptional flexibility in creating interesting, topical assignments. Instructors can opt to author new assignments, or select assignments from an online library that includes at least 24 assignments on neuroscience related topics (i.e., cognitive neuroscience, psychopharmacology, ethics, cellular and molecular neuroscience, visual neuroscience, signal transduction, clinical neuroscience, and neuroethology). Using a simple,

online interface, instructors provide assignment goals, guiding questions, essay word length (Fig. 1), and links to relevant source materials (e.g., lecture notes, primary research articles, and scholarly search engines). Next, instructors compose high, low, and mid-quality calibration essays (Fig. 2), set the evaluation parameters (ten style and content questions that can be answered in either yes/no or none/one/some categorizations; Fig. 3), and evaluate the calibration essays according to these criteria. Finally, instructors set the deadlines for text entry and peer review completion, and establish the grading criteria for the calibrations (the number of specific questions a student must answer correctly, and the acceptable deviation range for the holistic evaluation of the essays). Excellent technical advice is available through email and a list-serve, and formal workshops on authoring successful assignments are offered on a regular basis. Consult <http://cpr.molsci.ucla.edu> for more details about the program.

Sample Summary Assignment

Retrograde Bliss: Synaptic signaling of endocannabinoids

Although the drug cannabis has been used as a psychoactive substance for more than 4000 years, it has only been in the last decade that neuroscientists have begun to understand the drug's mechanism of action through studying endocannabinoids, endogenous substances that bind to the cannabinoid receptor. Your task is to write a 500-725 word summary of the article "The Brain's own Marijuana" by researchers Nicoll and Alger (*Scientific American*, 2004). Be sure to answer carefully and completely the following guiding questions:

What functions do the endocannabinoids serve?

Do the endocannabinoids fit the classic definition of neurotransmitters? What is the endocannabinoid mechanism of action at the synapse?

How might these compounds (or their synthetic analogues) be used therapeutically?

Figure 1. CPR assignment goals and guiding questions used in the introductory neuroscience course. This assignment included direct links to the review article and to the professor's lecture slides on neurotransmission.

Completing assignments

Before students begin any assignment, a mandatory, online interactive tutorial familiarizes students with using the CPR interface to complete each stage of the assignment. Completing CPR assignments requires three phases of student participation: 1) composing an essay, 2) calibrating reviews against standards, and 3) reviewing self and others. Students first compose and electronically enter their essays by a certain deadline; this can be done from any internet-ready computer.

Second, students evaluate three instructor-provided calibration essays holistically by numerically rating the essay from 1-10, and specifically by answering content and style questions set by the instructor (Fig. 3). After completing the three calibration reviews, students learn how their calibration evaluations compare with the instructor's evaluation (i.e., the number of style and content questions answered correctly, and the deviation from the instructor's holistic score for each calibration essay.) If

students do not “pass” a calibration evaluation, they have one more opportunity to revise their assessments before receiving their official scores on the calibration section. This second feedback in the calibration phase includes extensive comments written by the instructor to clarify student understanding of the judgment criteria.

Excerpts from Calibration Essays

High Quality

Endocannabinoids are similar to traditional neurotransmitters, yet several distinct features separate these compounds. A neurotransmitter is produced in the presynaptic terminal or in the soma, and is stored in vesicles within the neuron. Released from the presynaptic terminal, a neurotransmitter acts postsynaptically and is deactivated by transporter proteins or enzymes. In contrast, the postsynaptic cell releases endocannabinoids. They then travel “upstream” and bind with the CB1 receptors on the presynaptic cell, in a previously unknown method of communication called “retrograde signaling.” Once bound to the receptor, an endocannabinoid activates a G-protein, which generally initiates a decrease of GABA release. As the effects of marijuana use suggest, CB1 receptors are heavily distributed on GABA-releasing neurons throughout the hypothalamus, amygdala, neocortex, brain stem, basal ganglia and cerebellum.

Mid-quality

Marijuana makes some people hungry because it acts on the hypothalamus, an antagonist blocker for the endocannabinoid receptors could be made into a diet pill that really works. It is important to remember though that endocannabinoids are not actually neurotransmitters as they're traditionally defined. Endocannabinoids are released from all throughout the neuron but generally nearby the synapse. Neurotransmitters are released into the synaptic cleft. Vesicles store and recycle neurotransmitters. Endocannabinoids work with a g protein second messenger. Endocannabinoids are synthesized as they're released. Their not made in the axon or soma, and they're not stored in the vesicles. They can be thought of as a new school neurotransmitter, because these new compounds have actions very similar to a neurotransmitter, but are the next generation (Nicoll and Alger, 2004). Ultimately, they help modulate synaptic transmission like a metabotropic neurotransmitter.

Low Quality

Thus researchers are working across the world to discover the full functions of retrograde signaling and this fantastic form of neuron communication that allows neurons to do things that we previously thought were impossible and so never understood. It is obvious that the communication between the fornix and frontal lobe have many receptor sites for endocannabinoids. They act on neurons that release GABA which slows signaling down. Endocannabinoids are a newly discovered kind of neurotransmitter because they act on synapses, and work with a G-protein. Endocannabinoids modulate synapses and endocannabinoids are naturally produced. There are probably other things in the primate cortex that use the same type of exciting new communication so the implications go way past how we could use weed and endocannabinoids to do good things. Now that the scientists know that nerves can chat this way, they can try to use it in new theories to discover other chemicals and processes that will use this. Then we can use that knowledge to open other doors for more medical and neurological advancement.

Figure 2. Excerpts from the high-, medium-, and low-quality calibration essays. Before peer reviewing any essays, students were required to practice the evaluation process by reviewing three standardized ‘calibration’ essays on the same topic.

Thus, all students in the class receive the same extensive training in peer review prior to evaluating others’ essays. Next, the CPR program anonymously presents three essays for peer review. Students evaluate the essays according to the same criteria as the calibration essays and provide written justifications for each answer.

If a student does not perform satisfactorily on the calibrated reviews, then his or her peer reviews do not factor into the calculation of the peer-reviewed student text ratings. This provides an effective means for quality control in generating the text ratings. Finally, after reading six essays on the same topic, students evaluate their own essays using the evaluation parameters.

Evaluation Parameters

1. Is the main argument of the essay effectively conveyed in the thesis statement?
2. Does the essay address the guiding questions?
3. Are specific examples provided from both articles that provide support for the major argument of the essay?
4. When citing supporting evidence, does the essay include sufficient experimental detail and specific results?
5. Are there key words and/or concepts that are not well-defined in the essay?
6. Are there major factual errors in the essay?
7. Does the essay allow the reader to distinguish between proven results and implication and speculation?
8. Is the essay organized in a logical manner?
9. Are the sentences generally composed in a clear, concise manner?
10. Are there more than a few (1 or 2) spelling and grammar errors which show carelessness on the part of the author or detract from the overall quality of the essay?
11. Rate this essay on a scale of 1 (lowest) - 10 (highest).

Figure 3. Example evaluation criteria. These content (1-8) and style (9-10) questions address the quality of the essay at the global (main idea validity of argument), paragraph (transitions between and organization of paragraphs), and sentence (word choice, sentence structure, grammar and spelling) level. Question 11 asks the student to provide a holistic text rating for the essay. These questions are composed by the instructor and changeable with each assignment, but must be able to be answered in a yes/no or none/one/some format.

CPR Scoring

Immediately after the deadline for the completion of the peer and self reviews, students are able to access their peers’ narrative comments and receive a comprehensive score for the assignment based on four factors: essay quality, calibration review accuracy, peer review accuracy, and self review accuracy. Instructors determine weight of grading components; for both classes I weighted text 15% (5% each essay), the accuracy of three peer reviews quality 60%, the accuracy of the three calibration reviews 15% (5% each essay), and the accuracy of the self review 10%. The accuracy of these reviews is determined by correlation with other students’ evaluations. A scoring algorithm flags the instructor to any potential errors in analysis (e.g., essays that are reviewed by fewer than three peers, peer reviews that vary significantly in holistic evaluation, or essays that are reviewed by students who have done poorly in the calibration stage). Throughout all stages of the assignment, the instructor has access to all student work, including text entries, calibration and peer reviews. Instructors also know which students reviewed each essay, and are able to change essay scores and edit

assignment grades. The instructor can provide written commentary about grade changes, and has the option of making this logic visible to the students who peer reviewed the essay in question, or just to the author of the essay. These results are permanently stored in a user-friendly database on a UCLA server, and can be accessed by the instructor at any time.

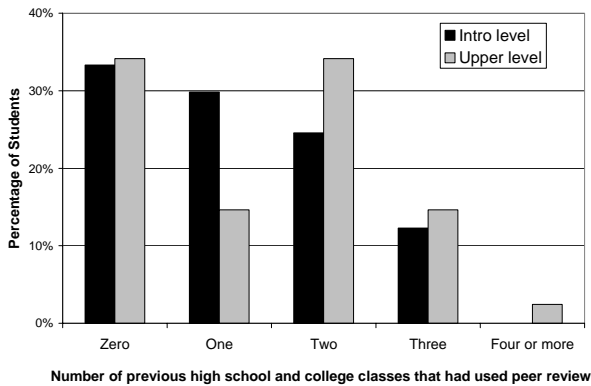


Figure 4. Previous student experience with peer review. Over one third of the members of both the introductory and the upper-level classes reported never using peer review in a high school or college setting.

Student Performance

The mean text ratings (generated by the weighted average of three anonymous peer reviews) and the mean deviations on calibrations, peer reviews, and self reviews are listed in Table 1. The CPR-calculated text ratings were within 0.5 points (on a ten-point scale) of the instructor's assessment 75% of the time. When the essays were rescored, the changes usually improved the students' text ratings, consistent with the finding that peer assessment of writing tends to be slightly lower than an instructor or teaching assistant's evaluation (Stefani, 1994). In both classes, the average deviation from the actual text rating was greatest in the calibration phase and least in the peer review phase (Table 1), suggesting that the calibration process did indeed function successfully as a training mechanism. However, the mean deviation for the self reviews, performed after the peer reviews, was greater than the mean deviation in peer review stage; students usually rated their own essays higher than their peers' assessment of the text. Similar to the findings of Falchikov and Goldfinch (2000), advanced students in the 300-level class did not perform any better at assessments than the beginner students in the introductory class, suggesting that peer review is not necessarily a skill served in the normal course of undergraduate academic training.

Student Evaluation of CPR

Student quantitative and qualitative assessments of the CPR program are summarized in Figure 5 and Table 2. Students self-reported that both their writing and peer review skills improved "somewhat" over the course of the semester (Fig. 5). Feedback on the utility of the peer

review process was mixed. Students reported that they "usually" provided helpful comments in peer reviews, whereas they only "sometimes" received helpful comments, and only "sometimes" felt like their essays were being evaluated fairly (Fig. 5). However, most all students agreed that, compared to other writing assignments of equal length, they spent much more time on CPR assignments. Not surprisingly, when asked to name their primary dislike of the CPR program, students cited "too much work" more often than any other answer (Table 2). However, students did report an increased sense of empathy for instructor workload.

The second most common complaint was the sense of being graded "subjectively" by their peers. The students, by virtue of reviewing the essays in a double-blind anonymous format, lacked the requisite information to 'subjectively' evaluate essays. Students occasionally did poor jobs at evaluating essays, but in these instances, their peer reviews were usually automatically discounted by the CPR program because these students also tended to fail the calibration section. The added motivation of being graded on the reliability of one's peer review further encouraged consistency and thoughtfulness in the peer review process. Nonetheless, students did cite 'grade-scheming' and attempts to 'second guess' the scoring algorithm as a major disadvantage of the program. For example, when the "acceptable" deviation for a peer review was ± 2 , some students tended to rate a very good (9 or 10 quality) essay as an 8 in order to maximize their acceptable deviation range.

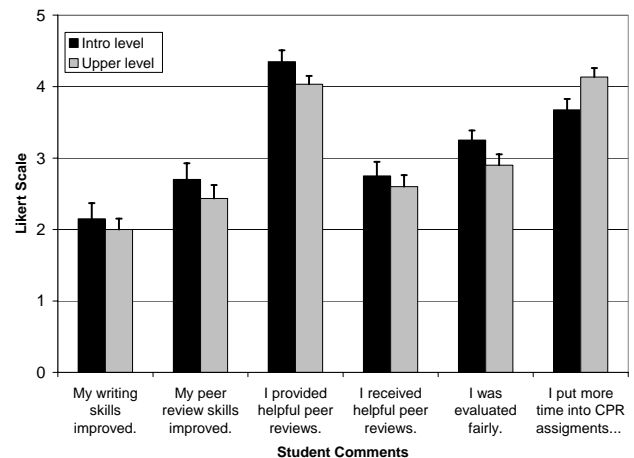


Figure 5. Student evaluation of the CPR program. At the end of the semester, students ranked on a 1-5 scale how much their writing and peer review skills were improved over the semester, with 1 = none and 5 = very much; how often they provided and received helpful peer reviews and how often they felt like their essays were evaluated fairly, with 1 = never and 5 = always; and how much time they put into CPR assignments, compared to other writing assignments of similar length and complexity, with 1 = much less and 5 = much more.

Importantly, students voiced very few negative comments about CPR being inaccessible, hard to understand, or technically challenging. Students reported feeling comfortable with the format after only one

assignment. Although more students stated that the online format was a deterrent (5) than an advantage (2), technical difficulty with the program itself was never mentioned as a concern. By far, most student confusion stemmed from ambiguities in the criteria used to evaluate essays (Fig. 2); specifically, students struggled to answer the evaluation questions with either yes or no answers. To address this concern, I offered full credit for a calibration review if a student holistically scored the essay within the acceptable range, but answered more of the content or style questions incorrectly than was permitted, only if the student was able to justify his or her answers with examples from the text. Another strategy to minimize confusion about the evaluation parameters is to spend time in class collectively evaluating a sample essay before students begin the first assignment, as peer assessments have been found to resemble instructor assessments most closely when judgments are based on well understood criteria (Falchikov and Goldfinch, 2000).

| Liked most about CPR | % Cited |
|--|---------|
| Interesting topics / Real world applications | 26% |
| Benefited from peer review process | 17% |
| Deepened my understanding of the material | 11.5% |
| Showed comprehension in non-test format | 10% |
| Appreciated opportunities to revise | 8.5% |
| Honed my editing skills | 6% |
| Preferred online accessibility | 3% |
| It forced more work out of peers | 1.5% |
| I preferred the anonymity | 1.5% |
| No response | 16% |
| Disliked most about CPR | % Cited |
| Peer review is too much work | 27% |
| Subjective grading by peers | 23% |
| Being graded on accuracy of peer reviews | 11% |
| Rigidity of the calibrations (Y/N format) | 10% |
| Reviews are too varied in usefulness | 8.5% |
| Online format | 7% |
| Grade scheming by peers | 4% |
| Rigid word limit | 4% |
| Unforgiving deadline | 3% |
| Too much emphasis on grammar | 1.5% |
| No response | 16% |

Table 2. Qualitative student evaluation of the CPR program. Students were asked to name what aspect they liked most and least about the CPR program in an anonymous survey. Answers from the introductory and upper-level classes were combined.

Instructor Evaluation of CPR

Despite student critiques of the program, I still recommend the CPR program as an excellent and free resource for incorporating more writing, peer review, and critical thinking into an undergraduate neuroscience curriculum. CPR's online data management system also has many

practical advantages to the traditional paper collection method of peer review. First, the random double-blind distribution of peer essays simplifies a very difficult organizational feat and guarantees student anonymity. Second, the saved database of peer reviews is a great source of information for writing recommendation letters and for calculating participation grades in large classes. Third, the online submissions process eliminates problems with lost essays or broken printers, and the password-protected website for the class allows students to print off specific articles for the assignments without the burden of going to the library reserves, or worrying about copyright violations.

The design of the CPR program is pedagogically sound. Although I did not specifically measure whether learning was improved in topics covered by CPR assignments, many students cited a deepened understanding of the material as their primary like of the CPR program (Fig. 5). Other studies have shown quantitative learning gains in topics presented in CPR assignments, as compared to those topics covered in didactic lectures or in active learning formats (Pelaez, 2002). Perhaps one of the greatest advantages of CPR is the availability of immediate feedback on assignments. Learning is generally improved by detailed and timely feedback on student work (Brown et al., 1995), and using CPR, students received three detailed peer assessments of their work within five minutes of the deadline for completing the peer and self reviews. Furthermore, a quick online review of the essays allowed me to identify and address common content misconceptions in class and common errors in writing in the next set of calibration essays.

My two primary critiques of the CPR program are that 1) there is no process for revision built into the CPR program, and 2) the peer review component of the grade is determined by the quality of the students' quantitative, but not qualitative, feedback. Thus, students can receive full credit for reviews that are sloppy, pithy, or cruel in content, but within the allotted margin of error. To address the first issue, I offered students an opportunity to revise their essays for extra credit (a maximum of 40% of the difference between their grade and a perfect score). Sixty-three percent of students in the introductory class opted to revise at least one essay, whereas only 40% of students in the upper level class chose to revise an essay (Table 1). This added revision process allowed me to provide students with more personal attention, and conversations with students about what peer advice to adopt and what suggestions to ignore were generally very fruitful. To address the second issue, I advised students that the quality of their written peer review comments would factor heavily into their class participation grade, 5% of their overall grade for the course.

Although CPR was designed as a timesaving device for instructors, the workload associated with introducing the program and discussing the evaluation parameters using example essays (1.5 class periods), creating the assignments and the calibration essays (8-10 hours), and re-evaluating text ratings and contested calibration and

peer reviews (4-8 hours) is considerable. Of course, repeating old assignments, using an assignment from the CPR library, or hiring a qualified student assistant to compose the calibration essays would significantly expedite this process. Nonetheless, CPR still requires less time than critically responding to 30+ student essays by hand, and provides far more direct experience for the students in abstract reasoning and peer review.

Results of the student evaluations indicate that CPR fostered a multi-dimensional comprehension of the course material while teaching traditionally underserved academic skills: science writing and peer review. Over the course of the semester, I observed that students' purpose for writing shifted from writing for the professor, to writing a clear argument for a general audience. Indeed, one of the downfalls of traditional instructor review is that comments by instructors detract students' attention from their own intentions in writing, and focus that attention instead on "the teachers' purpose in commenting" (Sommers, 1982). CPR certainly requires more student engagement and autonomy in the writing and review process than traditional assignments.

REFERENCES

- Bean JC (1996) *Engaging ideas: The professor's guide to integrating writing, critical thinking, and active learning in the classroom*. San Francisco, CA: Jossey-Bass.
- Bloom BS, ed. (1956) *Taxonomy of educational objectives, cognitive domain*. New York, NY: Longmans, Green and Company.
- Brown S, Race P, Rust C (1995) Using and experiencing assessment, in Knight P (ed.) *Assessment for Learning in Higher Education*, London: Kogan Page/SEDA.
- Chapman OL (1999) Calibrated peer review™, an overview. <http://cpr.molsci.ucla.edu>.
- Committee on Undergraduate Science Education (1997) *Science Teaching Reconsidered: A handbook*. Washington, DC: National Academy Press.
- Falchikov N, Goldfinch J (2000) Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Rev Educ Res* 70:287-322.
- Nicoll RA, Alger BE (2004) The brain's own marijuana. *Sci American* 291:69-75.
- Peleaz N (2002) Problem-based writing with peer review improves academic performance in physiology. *Adv Physiol Educ* 26:174-84.
- Rao SP, DiCarlo SE (2000) Peer instruction improves performance on quizzes. *Adv Physiol Educ* 24: 51-5.
- Russell AA (2004) What Works- A pedagogy: Calibrated peer review. Project Kaleidoscope: What works, what matters, what lasts 4: http://www.pkal.org/template2.cfm?c_id=1317
- Sommers N (1982) Responding to student writing. *College Composition and Communication* 33:148-56.
- Stefani LAJ (1994) Peer, Self and tutor assessment: relative reliabilities. *Studies in Higher Education* 19:69-75.

Received August 01 2005; revised October 07, 2005; accepted October 16, 2005

The author thanks Tim Su for his excellent work in maintaining the CPR website, Arlene Russell for training and advice, Keith Hengen for composing the calibration essays, and Ruth Benca, MD, PhD, for initial mentorship in teaching writing and peer review.

Address correspondence to: Dr. J. Roxanne Prichard, Psychology Department, Bates College, 2 Andrews Rd., Lewiston, ME 04240. Email: jprichar@bates.edu.