Supporting scientific writing and evaluation in a conceptual physics course with Calibrated Peer Review

Edward Price[*], Fred Goldberg[^], Scott Patterson[^], and Paul Heft[*]

[*]*Department of Physics, California State University, San Marcos, CA 92096*
[^]*Department of Physics, San Diego State University, San Diego, CA 92182*

Abstract. Writing tasks are one way students can apply science concepts, yet evaluating students' writing can be difficult in large classes. With the web-based Calibrated Peer Review* (CPR) system, students submit written work and evaluate each other. Students write a response to a prompt, read and evaluate responses prepared by the curriculum developers, and receive feedback on their evaluations, allowing students to "calibrate" their evaluation skills. Students then evaluate their peers' work and their own work. We have used CPR for two semesters in conceptual physics courses with enrollments of ~100 students. By independently assessing students' responses, we evaluated the CPR calibration process and compared students' peer reviews with expert evaluations. Students' scores on their essays correlate with our independent evaluations. This poster describes these findings and our experiences with implementing CPR assignments.

*Calibrated Peer Review, http://cpr.molsci.ucla.edu/

Keywords: conceptual physics, scientific writing, peer-review, CPR
PACS: 01.40.G-, 01.40.Di, 01.40.Fk

INTRODUCTION

Learning Physics (LEP) is a new inquiry-based, conceptual physics curriculum that is suitable for a large lecture hall environment. LEP is part of a family of related curricula: Learning Physical Science (LEPS) [1], Physical Science and Everyday Thinking (PSET) [2], and Physics and Everyday Thinking (PET) [3]. PET and PSET are intended for small, discussion-based courses, while LEPS and LEP are designed for large enrollment classes. LEP is a one-semester curriculum with a student-oriented pedagogy designed to enable students to develop a deep understanding of the conceptual themes of conservation of energy and Newton's laws, as well as other topics. LEP is also intended to enable students to develop an understanding of important aspects of scientific thinking and the nature of science (NOS).

The LEPS curriculum was adapted to large-enrollment settings from PSET [1]. While we were fully able to adapt many pedagogical features, LEPS lacked or only partially included two important pedagogical features: opportunities for hands-on exploration of phenomena (other than the optional labs), and construction and evaluation of scientific explanations. In developing LEP, we are attempting to address these issues; this paper focuses on the science practice of constructing and evaluating explanations. To accomplish this within the context of a large enrollment course, LEP includes five tasks that students complete using Calibrated Peer Review (CPR) [4]. CPR is a web-based tool that supports students' construction and evaluation of explanations. The system incudes a peer review process with a training component to prepare students for reviewing. A student's grade on a CPR task is based on, among other things, her peers' reviews of her work; the instructor does not have to grade the students' essays. CPR has previously been used in a bioengineering lab course where students make technical posters [5]. During fall 2011 and spring 2012, we implemented CPR tasks in pilot versions of the LEP curriculum. This paper reports some initial results from those tests, and addresses the research question*, "What is the validity of the peer evaluation process in CPR?"*

BACKGROUND – CPR SYSTEM

In PET and PSET, students develop and practice the construction and evaluation of explanations of phenomena. However, in large enrollment courses, the instructor may not be able to evaluate and provide feedback on written work. A goal in the development of the LEP curriculum is to address this issue with CPR-based tasks. CPR provides scaffolding for students to learn how to evaluate and construct explanations, but without an unreasonable grading load on the instructor.

CPR tasks include three stages, as illustrated in Figure 1. In the text entry stage, students explore background material and write a response, which is submitted through the CPR website. In LEP, students might read a description of a phenomenon or watch a video of an experiment, and create an explanation using text and drawings. After the text entry stage, the calibration and review stage begins. Students read a set of "calibration" essays prepared by the curriculum developers to represent low, medium, and high quality responses to the prompt. The students evaluate the calibration essays according to specific questions established by the developers and give each essay an overall score. The developers also supply answers and feedback for the evaluation questions, along with an overall score for each calibration essay. The CPR system then shows students how their evaluations compare to those provided by the developers, including feedback on the evaluation questions, which allows students to "calibrate" their evaluation skills.

After a student completes the calibration stage, she is assigned a Reviewer Competency Index (RCI). As described by the CPR documentation, "the RCI is based on two criteria: the number of calibrations passed and the average deviation of the last three ratings. The average deviation of ratings is the average of the absolute deviations between the calibration key ratings, determined by an assignment author, and the student rating for the three calibration essays." A calibration is "passed" if a student answers a set number of the evaluation questions correctly and gives an overall evaluation score that agrees with the developer's score (within a set range). If students do not pass a calibration, they can repeat it once. In the spring 2012 pilot test of LEP, students passed a calibration if they answered at least 9/10 evaluation questions correctly.

The RCI is an integer between 0 and 6. Students with low average deviations who passed all three calibrations receive an RCI of 6. Students with larger average deviations or fewer passing calibrations receive lower RCIs. The CPR system uses the RCI to determine a student's competence as a reviewer.

Students next evaluate the work of several of their classmates and their own work. The score a student receives on her text is based on a weighted average of the scores given to her by three of her peers, with the RCI used as the weighting factor. Thus, scores from students with low RCIs are counted less than scores from students with high RCIs [6]. The resulting score is called the Average Weighted Text Rating (AWTR).

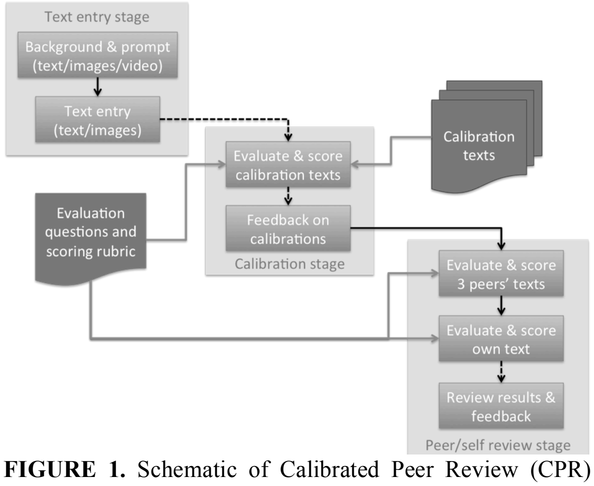

Text entry stage
Background & prompt (text/images/video)
Text entry (text/images)
Calibration texts
Evaluate & score calibration texts
Feedback on calibrations
Calibration stage
Evaluation questions and scoring rubric
Evaluate & score 3 peers' texts
Evaluate & score own text
Review results & feedback
Peer/self review stage

FIGURE 1. Schematic of Calibrated Peer Review (CPR) system. CPR uses peer review with a training, or calibration, stage to prepare students for reviewing.

Finally, in the results stage, students can review how their evaluations compared to those of other students (three students evaluate each explanation). Students also review other students' evaluations of their own explanation. Students are scored based on the quality of their text (based on their classmates' evaluation), their calibrations (were they consistent with the curriculum developers?), their peer reviews (were their reviews consistent with their classmates' reviews?), and the quality of their self-assessment (was their evaluation of their own work consistent with their classmates' evaluations of it?).

The system flags problematic results for review by the instructor; a problem may result from reviews by students with very low RCI scores (2 or less), or large deviations between reviewers' scores of students work.

USE OF CPR IN LEP

CPR assignments were used in a fall 2011 pilot implementation of LEP, and then substantially revised. In spring 2012, students were assigned five CPR tasks, an initial practice task and one task for each major unit in LEP (interactions and energy, light, forces and motion, and a model of magnetism).

The Unit 1 (U1) task required students to construct energy transfer diagrams for a chain of interacting objects and determine the energy efficiency. The Unit

1. On a single sheet of paper draw two iron nails. Label one "unmagnetized nail" and the other "magnetized nail." Using the Alignment Model, draw the entities inside the unmagnetized nail. Next, draw the entities inside the magnetized nail, and label the poles (taking into account the situation described above). Upload a picture of your diagrams.

2. In the 1st paragraph describe how you have drawn your diagram for the unmagnetized nail; that is, what is your diagram trying to show. Also explain why the nail is unmagnetized; that is, why it produces no magnetic effects in the region outside the nail. [You need to use the Alignment Model.]

3. In the 2nd paragraph explain how you know, based on the evidence provided, whether the tip end of the magnetized nail is a NP or a SP. [You need to state and use the appropriate law.]

4. In the 3rd paragraph explain how you know which end of the magnet, its NP or its SP, was used to slide across the nail from head to tip. [You need to use the Alignment Model and state and use the appropriate law.]

5. In the 4th paragraph explain why hammering the magnetized nail caused it to become unmagnetized. Begin by describing your drawing for the magnetized nail, and then explain what happened when the nail was hammered. [You need to use the Alignment Model.]

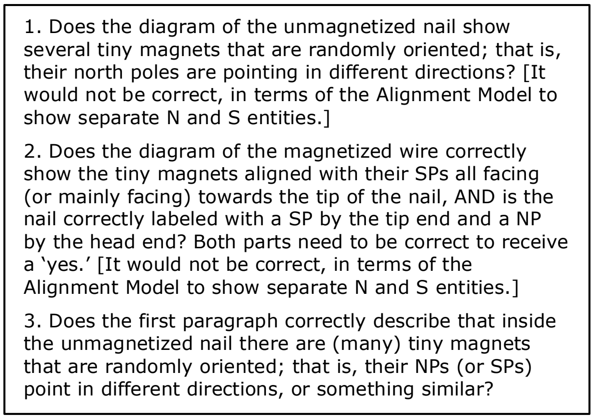
FIGURE 2. Example CPR assignment from the model of magnetism unit.

2 and 3 CPR tasks were more problem-solving oriented and are not discussed in this paper. Figure 2 shows the writing prompt for the Unit 4 (U4) CPR task [7]. U4 focuses, in part, on a Model of Magnetism.

For the CPR task, students read a description and watched a video of an initially unmagnetized iron nail being magnetized, then hit with a hammer so that it becomes unmagnetized. At each step the nail is slid towards a compass, along its E-W axis. The writing prompt instructs students to use the model of magnetism developed in class to draw diagrams of the nail and write an explanation for the observed behavior. Figure 3 shows the evaluation questions for the U4 CPR task. Students were instructed to score the essay from 0-10 based on the number of evaluation questions to which they answered “yes.”

The fall 2011 pilot test led to a number of refinements in spring 2012, including 1) more specific writing prompts and 2) more focused/specific evaluation questions. As a result of these changes, in the spring 2012 implementation more students passed the calibrations, average RCI scores increased, fewer results were flagged as problematic, and there were far fewer student complaints.

Near the end of the spring 2012 semester, interviews were conducted with six students (two each with high, middle, and low quiz grades). Students were asked about their perception of the CPR system (how it was used and how the students thought it went), how helpful it was for learning and/or getting a good grade, and their understanding of the purpose for using CPR in the course. Students’ responses were mixed. Four described as valuable the process of creating an explanation and evaluating other’s work. Students also expressed frustration with the multiple parts and complexity of the CPR assignment (four students). Finally, two students described frustration at part of their grade being determined by other students.



1. Does the diagram of the unmagnetized nail show several tiny magnets that are randomly oriented; that is, their north poles are pointing in different directions? [It would not be correct, in terms of the Alignment Model to show separate N and S entities.]

2. Does the diagram of the magnetized wire correctly show the tiny magnets aligned with their SPs all facing (or mainly facing) towards the tip of the nail, AND is the nail correctly labeled with a SP by the tip end and a NP by the head end? Both parts need to be correct to receive a ‘yes.’ [It would not be correct, in terms of the Alignment Model to show separate N and S entities.]

3. Does the first paragraph correctly describe that inside the unmagnetized nail there are (many) tiny magnets that are randomly oriented; that is, their NPs (or SPs) point in different directions, or something similar?

FIGURE 3. The first three (of ten) “yes” or “no” evaluation questions for the model of magnetism unit CPR task. The score on the task is number of “yes” responses to the evaluation questions.

METHODS

We analyzed the Unit 1 and Unit 4 tasks used in spring 2012. There were 108 students in the course. For each task we scored the students’ texts using the same evaluation questions and procedure used by their peers. We refer to this researcher-generated score as a student’s R-score. With R-scores for all students, we could then compare the scores students received from their peers and their R-scores. In particular, we compared AWTR, which is the CPR system’s best determination of student’s text score, with R-scores.

We also used the R-scores to evaluate the quality of students’ reviews of their peers. To do this, we computed a Peer-Reviewer Competency Index (P-RCI) in analogy to the RCI. The RCI compares a student’s evaluation to the developer’s evaluation of the calibration essays; this is possible since the developer prepares the calibration essays in advance. Similarly, the P-RCI compares the student’s evaluation of their peers to the R-score of their peers. Consider student A, who reviews and scores students B, C, and D. We compare student A’s score for B with B’s R-score, etc. Of course, this type of comparison is only possible because we have independently scored each student’s work; typically, an expert scoring of each student’s work is not available (which is a large motivation for peer reviewing in the first place).

We calculated average values and standard error of the mean (SEM) for the AWTR, R-scores, RCI, and P-RCI for the U1 and U4 CPR tasks. Values for these scores are not normally distributed. Further, the R-score, RCI, and P-RCI take on integer values. For this reason, the SEM should be interpreted with caution. We compared R-scores with AWTR and RCI with P-RCI using a two-tailed Wilcoxon test, which is a non-parametric significance test appropriate for paired, non-Gaussian data.

RESULTS AND DISCUSSION

The averages and SEM for the AWTR, R-scores, RCI, and P-RCI are shown in Table 1. We first compare R-score and AWTR. The average AWTR was greater than the average R-score for both tasks. The difference is statistically significant for the U1 task, but not for U4, as indicated in Table 2. A Pearson linear correlation test yielded R^2 = 0.67 (significant with $p<0.001$) for U1, and R^2 = 0.68 (significant with $p<0.001$) for U4.

TABLE 1. Average & standard error of the mean (SEM).

	U1 avg	U1 SEM	U4 avg	U4 SEM
R-score	8.3	0.15	8.7	0.21
AWTR	8.7	0.12	8.9	0.16
RCI	5.2	0.11	4.8	0.13
P-RCI	4.4	0.13	4.5	0.14

A student's grade on a CPR task depends on many factors, including her reviews of the calibration essays and her peers' explanations. However, the AWTR is the systems' best measure, based on her peers' evaluations, of her writing. In the U1 and U4 tasks, based on the degree of correlation and small difference between averages for the R-score and AWTR, CPR's peer-reviewing process leads to valid AWTR scores for most students.

We next compare P-RCI with RCI. This gives a measure of the validity of the RCI score: what is the relationship between the CPR system's evaluation of a student's competence as a reviewer and her actual performance on peer reviews? As shown in Table 1, the average RCI score is higher than the P-RCI score for both tasks, although the difference is only significant for U1 (see Table 2). For the U4 task, we compared the RCI and P-RCI scores, by grouping scores into a low range (0-3) and high range (4-6). Of the 108 students, 4% were low-low (RCI-PRCI) and 70% were high-high. In these cases, the RCI accurately indicated the competence of their peer reviewing (within the high/low designation). In contrast, 8% were low-high and 18% were high-low. The low-high students are false negatives: they were better peer reviewers than expected based on their low RCIs. The high-low students are the problematic case of false positives; they were inaccurate peer reviewers (hence a low P-RCI) but are weighted heavily by the system (due to a high RCI). Because each student is reviewed by three others, the 18% high-low reviewers can have a sizeable impact on the CPR process.

The P-RCI might differ from the RCI because the activities are different (either because the student essays differ from the calibration essays in an important way, or because the students treat the two tasks differently, *e.g.*, they are unwilling to score their peers harshly), or because students learn during the calibration and/or peer reviewing process.

Finally, we consider the possibility of change during the semester. Students complete the U1 and U4 tasks near the start and end of the semester, respectively. Because the tasks involve different concepts, prompts, and evaluation questions, a direct comparison between U1 and U4 scores is not meaningful. Instead, we compare differences for each task. For both tasks, the average AWTR is higher than the average R-score, and the average P-RCI is lower than the average RCI. However, neither difference is statistically significant for U4. This could be evidence for improvement in students' ability to write and evaluate explanations. However, an increased familiarity with the CPR system and level of understanding of the unit content could also be factors.

TABLE 2. P-values for Wilcoxon matched pairs test

	U1	U4
R-score vs AWTR	<0.001	0.375
RCI vs P-RCI	<0.001	0.215

CONCLUSIONS

We used CPR to support writing and peer evaluation in a large enrollment, conceptual physics course. Students' scores based on peer evaluations compare well to scores given by researchers. The calibration process employed by CPR provides a reasonable estimation of students' competence as a reviewer, although a sizeable fraction of students perform reviews below the system's expectation. Finally, there is some evidence for development in students' ability to write and evaluate explanations.

ACKNOWLEDGMENTS

The authors would like to acknowledge Mike McKean, Melissa Dancy, CSUSM's IITS, and NSF DUE grants 0717791 and 1044172.

REFERENCES

1. F. Goldberg, E. Price, S. Robinson, D. Boyd-Harlow, and M. McKean. *Phys. Rev. ST Phys. Educ. Res.* **8**, 010121, 2012.
2. F. Goldberg, S. Robinson, R. Kruse, N. Thompson, and V. Otero, *Physical Science and Everyday Thinking, 2nd ed.* It's About Time: Armonk, NY, 2009.
3. F. Goldberg, S. Robinson, and V. Otero, *Physics and Everyday Thinking,* It's About Time: Armonk, NY, 2005.
4. Calibrated Peer Review, http://cpr.molsci.ucla.edu/ accessed July 1, 2012.
5. T. Volz, & A. Saterbak. Students' strengths and weaknesses in evaluating technical arguments as revealed through implementing Calibrated Peer Review in a bioengineering laboratory. *Across the Disciplines*, **6,** Special issue on Writing Technologies and Writing Across the Curriculum, 2009. http://wac.colostate.edu/atd/technologies/volz_saterbak.cfm
6. Complete details are in the CPR documentation, available at http://cpr.molsci.ucla.edu/Downloads.aspx
7. A full description of the magnetism task is available at http://faculty.csusm.edu/price/LEP/CPR_example.html