Validity of peer grading using Calibrated Peer Review in a guided-inquiry, conceptual physics course

Edward Price,^{1,*} Fred Goldberg,² Steve Robinson,³ and Michael McKean²

¹*Department of Physics, California State University, San Marcos, California 92096, USA*

²*Department of Physics, SDSU, San Diego, CA 92182, USA*

³*Department of Physics, TTU, Cookeville, TN 38505, USA*

(Received 22 March 2016; published 22 December 2016)

Constructing and evaluating explanations are important science practices, but in large classes it can be difficult to effectively engage students in these practices and provide feedback. Peer review and grading are scalable instructional approaches that address these concerns, but which raise questions about the validity of the peer grading process. Calibrated Peer Review (CPR) is a web-based system that scaffolds peer evaluation through a “calibration” process where students evaluate sample responses and receive feedback on their evaluations before evaluating their peers. Guided by an activity theory framework, we developed, implemented, and evaluated CPR-based tasks in guided-inquiry, conceptual physics courses for future teachers and general education students. The tasks were developed through iterative testing and revision. Effective tasks had specific and directed prompts and evaluation instructions. Using these tasks, over 350 students at three universities constructed explanations or analyzed physical phenomena, and evaluated their peers’ work. By independently assessing students’ responses, we evaluated the CPR calibration process and compared students’ peer reviews with expert evaluations. On the tasks analyzed, peer scores were equivalent to our independent evaluations. On a written explanation item included on the final exam, students in the courses using CPR outperformed students in similar courses using traditional writing assignments without a peer evaluation element. Our research demonstrates that CPR can be an effective way to explicitly include the science practices of constructing and evaluating explanations into large classes without placing a significant burden on the instructor.

DOI: [10.1103/PhysRevPhysEducRes.12.020145](https://doi.org/10.1103/PhysRevPhysEducRes.12.020145)

I. INTRODUCTION, BACKGROUND, AND CONTEXT

Constructing and evaluating explanations are important science practices, and developing students’ capabilities with these practices is a goal for many physics educators and physics education researchers. However, in many classes, especially classes with large enrollments, practical concerns make it difficult to effectively engage students in constructing and evaluating explanations, or to provide students with feedback on their efforts [1]. Peer review and grading is a scalable instructional approach that addresses these concerns and can be implemented in face-to-face or online courses at scale [2–5]. However, the use of peer review and grading raises questions about the appropriateness of students grading each other. This article explores those questions in the context of Calibrated Peer Review (CPR) [6], a web-based system that facilitates peer review of written

work. We used CPR in guided-inquiry, conceptual physics courses for preservice elementary teachers and general education students. This paper describes our use of CPR to engage students in the science practices of constructing and evaluating explanations of natural phenomenon, and evaluates the validity of the peer evaluation process in CPR.

A. Explanation, evaluation, and communication as science practices

In addition to content-related learning outcomes, physics educators and physics education researchers have articulated learning outcomes related to scientific practices or scientific abilities [7,8]. Argumentation, explanation, and communication are included among these practices, which is consistent with the science education community’s attention to writing, explanation, and argumentation as a core practice of science [9,10]. Calls to include science practices in science courses for nonscience majors are consistent with long-standing calls for science literacy [11].

Recently, and with important implications for the preparation of future teachers, the National Research Council’s *Framework for K-12 Science Education* [12] and the subsequent next generation science standards (NGSS) [13] emphasize the integration of science practices and science content. The framework describes eight science and

*Corresponding author.
eprice@csusm.edu

Published by the American Physical Society under the terms of the Creative Commons Attribution 3.0 License. Further distribution of this work must maintain attribution to the author(s) and the published article’s title, journal citation, and DOI.

engineering practices, including constructing explanations (for science) and designing solutions (for engineering); engaging in argument from evidence; and obtaining, evaluating, and communicating information. These three practices provide an explicit emphasis on explanation, evaluation, and communication. While the framework and NGSS are intended for K-12 science education, they have implications for university science instruction. Courses that prepare future teachers should be consistent with the framework and the integration of content and practices. The context of the present study includes such courses.

B. Peer evaluation

Previous researchers have explored the use, validity, and educational impact of peer evaluation, or grading, of written work [14–16]. Sadler and Good list four motivations or benefits from peer and self evaluation: logistical—the grading burden for the teacher can be reduced and feedback to students can be quicker and more detailed; pedagogical—students can learn from the process; metacognitive—students can develop increased skills of self-evaluation, judgment, and reflection; and affective—students may develop more positive attitudes towards assessment and feedback.

In general, research indicates that peer review provides a valid student evaluation (compared to review by an expert) and promotes learning [4,5,14,15,17–22]. Harris describes using peer-assessed laboratory reports in large biology classes, and reports excellent correlation between peer and expert scoring [18]. Walvoord reports on use of CPR in a zoology course [23], comparing expert scores to peer scores on 20 essays (representing approximately one-third of students in the course), and finding no significant difference for 3 out of 4 assignments. Sadler and Good trained students with a scoring rubric and report a very high correlation between peer-assigned grades and teacher-assigned grades, though they note some patterns of bias when students assigned grades [15]. Freeman reports on the use of peer assessment of written practice exams in a large introductory biology course, and finds that students were significantly easier graders than an expert, but that the student grading was “good enough to use” in their context (practice exams representing 11%–15% of the points in the course) [17]. Freeman also notes that students were less reliable peer graders of higher-order questions involving application, analysis, synthesis, and evaluation. Douglas *et al.*, used peer evaluation of video-based lab reports in blended and online courses [4,5]. Students created video lab reports, received training on evaluating the videos, and then peer reviewed other students’ videos. Douglas *et al.* found better agreement between student and instructor evaluation scores after having students evaluate “practice videos” and compare their evaluation to the instructor’s.

Peer evaluation is conducted in many ways, from having students directly exchange and score work to complex

online systems. The use of scaffolds such as scoring rubrics, training in the evaluation process, and expert feedback varies. The CPR system provides considerable structure and scaffolding for the peer evaluation task, and also manages the logistics of random, anonymous reviews. A number of published reports focus on the use of CPR in science classes and perceptions of its use [23–25]. Margerum used CPR in a general chemistry course [24]. Likkell used CPR in a large, introductory astronomy course and found that students using the CPR system for essay assignments reported an increase in their ability to accurately assess what they have written, compared to no change reported by students who wrote essays but did not use CPR [25]. In this work, we used CPR to engage preservice teachers and general education students in constructing and evaluating writing explanations for physics phenomena.

C. Calibrated Peer Review

CPR is a web-based tool that provides scaffolded support for student writing and evaluation, and utilizes peer grading instead of grading by the instructor. With CPR, graded writing assignments can be included in larger courses where the instructor grading time does not scale with student enrollment. The system includes a training component prior to the peer review process, in order to prepare students for reviewing. CPR was developed at the University of California, Los Angeles and has been used in a variety of courses [26].

The CPR system separates the curriculum developer and instructor roles. Developers can write assignments that are available in a central library. Individual instructors can then download relevant assignments to a course account, set deadlines and scoring criteria, and manage student accounts. Developing a CPR task includes creating background material and a prompt, writing a set of example responses, and creating questions that can be used for evaluating responses to the prompt. The example responses represent low, medium, and high quality work and are referred to as “calibration” responses in the CPR system. For each example response, the developers prepare answers to the evaluation questions along with feedback and an overall score; these materials show how an expert would perform an evaluation of the example response. These materials are all prepared in advance by a developer who may not be the same as the instructor in the course. In our use of CPR, one of us (F. G.) developed most of the CPR tasks, while three of us (E. P., F. G., and S. R.) used them in courses we taught.

Students perform a CPR task in three stages, as illustrated in Fig. 1. The stages are sequential; each stage has a deadline, then the next stage opens. In the text entry stage, students explore the background material and respond to the prompt via the CPR web interface. (The term “text entry” is somewhat misleading, because students can also

FIG. 1. Schematic of the Calibrated Peer Review process. CPR uses peer review with a training, or calibration, stage to prepare students for reviewing their peers' work. Students cannot proceed between stages until a deadline passes (indicated by dashed lines), but students can work within the stage at their own pace (indicated by solid lines). When appropriate, the system provides background material, calibration texts, evaluation questions, and other students' texts.

upload files, including images.) In our CPR assignments, 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 closes, the calibration and review stage begins. Students review a calibration "text," answer the evaluation questions, and give an overall score. The calibration texts are the sample responses prepared by the assignment developer. After a student has reviewed all three calibration texts, the CPR system shows her a comparison between her evaluations and the developers', including feedback on the evaluation questions. The student can repeat a review of a calibration response if needed. This training process allows students to calibrate their evaluation skills, and allows the CPR system to generate a rating of the student's performance as a reviewer.

After the calibration stage, students begin the review stage, when they anonymously evaluate the work of three (also anonymous) classmates and then their own work. For these evaluations, the students use the same evaluation questions they practiced with during the calibration stage. Of course, unlike the calibration responses, when students evaluate each other's work, the expert evaluation is unknown. The score assigned to a student response is the weighted average of the scores given to her by her three peers, with the weighting based on the reviewer's performance on the calibration stage. Thus, peer reviews from students who were inexperienced evaluators of the calibration responses are counted less than scores from students who were expert evaluators of the calibration responses. The

resulting score is called the average weighted text rating (AWTR). After students review their peers' work, they perform a self evaluation on their own response using the same procedures.

Finally, in the results stage, the student can review how her reviews compared to those of the two other students who evaluated the same peer. The student can also review her peers' evaluations of her own explanation. A student's score on the assignment is based on the quality of her text (the AWTR from her classmates), her calibrations (to what degree were they consistent with the curriculum developers?), her peer reviews (to what degree was her review of each text consistent with her classmates' reviews of the same text?), and the quality of her self-assessment (to what degree was her evaluation of her own work consistent with her classmates' evaluations of it?). When setting up the assignment, the instructor determines how much each component contributes to the overall grade for the task. The system flags problematic results for review by the instructor; a problem may result from reviews by students with very low calibration performances, or large differences between reviewers' scores of students work.

D. Instructional context—the learning physics curriculum

The context for the present study is a one-semester, inquiry-based, conceptual curriculum intended for preservice elementary teachers and other nonscience majors, which we call learning physics (LEP) [27]. LEP is suitable for classes with large enrollments or lecture-style rooms. LEP was developed using design principles based on research on science learning [28], including an understanding of learning as a complex process involving prior knowledge, behavioral norms, and interactions with tools and peers. The forerunner to LEP was physics and everyday thinking (PET), which is similar to LEP but was designed for smaller enrollment courses taught in laboratory- or studio-style rooms [29,30]. Other versions were also developed [31,32], and most recently the entire suite of materials have been updated and explicitly aligned with the *Next Generation Science Standards* [33]. As described in Sec. III A, PET courses were used as a comparison group in a portion of this study. While PET and LEP share similar curricular goals and design principles, the differences in the intended learning environments require differences in the curricula. A complete comparison is in Ref. [28]; here, we focus on how students engage in science practices related to constructing and evaluating written explanations for physical phenomena. PET students construct and evaluate written explanations in class, during small group work and whole class discussions. PET students also construct written explanations in homework assignments. Because of the difficulty in providing feedback in a large class, LEP students do not construct their own explanations in class. Instead, they are asked to evaluate sample explanations

provided by the instructor. Furthermore, given the large course enrollments, we did not assign written explanations for homework in LEP (as is done in PET). Instead, in LEP courses, we gave CPR-based homework assignments that included structured explanation and evaluation tasks.

II. THEORETICAL FRAMEWORK AND RESEARCH QUESTIONS

Our use and analysis of peer evaluation and CPR is guided by activity theory (AT) [34–37], which provides a framework for exploring the complex relationships between individuals, groups, and artifacts or tools. AT locates a subject (such as a physics student) within a community of people (other students, the instructor) sharing the same object (passing a class or learning a physics concept). The subject's actions are shaped by participation in the community and mediating tools. Rules and norms (implicit and explicit) prescribe how to go about the activity, answering the question, “How do things work here?” Roles, or a division of labor, describe who does what. These elements of the activity system all interact in a complex way.

Tools play a mediating role and shape the likelihood of possible actions. The concept of affordances helps when thinking about tools. Following Norman, we use affordances in the sense of “perceived affordances” as “the perceived and actual properties of the thing... that determine just how the thing could possibly be used” [38]. For example, a computer-based motion sensor affords the collection and graphing of data, making it easy to create and investigate graphical representations of motion. Tools also impose constraints; the motion sensor may only detect motion in one dimension. In AT, tools mediate activity; they can structure and reorganize action, change roles and norms, and enable new possibilities.

From this perspective, successful use of peer review by instructor and students will depend on consistent alignment between the tools, roles, and rules in the activity. Figure 2 represents these components and their interrelatedness. The CPR system is a tool that mediates students' engagement in constructing and evaluating explanations. More than just a way for students to submit an assignment, the CPR system has features and affordances designed to support the calibration and peer review processes. Implementing these components is prominent and natural in CPR, but would be much more difficult to do using a “paper and pencil” system. Introducing peer review restructures the roles in the classroom: students become evaluators in addition to generators of responses; the instructor becomes a facilitator of the process rather than an evaluator; and the role of materials developer is separated from the role of instructor. The CPR system supports students in their new role by providing mechanisms for training and feedback through the calibration process, utilizing the materials provided by the curriculum developers. Activity theory suggests that this restructuring of roles needs to be supported through the rules and norms of the classroom, which also have a mediating role. For instance, the instructor could foster supportive norms by emphasizing the value of evaluation skills and the importance of peer review in science. Supportive formal rules could include course grading policies that give credit for the CPR tasks. The CPR system supports establishment of these rules. For instance, the instructor can assign credit for completing the calibration and peer review stages of the assignment; and the system gives less weight to peer evaluations from students who did not perform the calibrations well. With this framing in mind, our research questions focus on the shift in students' roles to include that of evaluator: how can CPR, as a mediating tool, support this change, what changes are needed in other elements of the activity system to support

FIG. 2. The elements of the activity system and their interrelations.

this change, and how effective are students in this new role? Specific questions include the following:

- (1) What are the characteristics of successful CPR-based assignments in a conceptual physics courses? How specific do the prompts and evaluation questions need to be? What type of support do students need to successfully complete the assignments?
- (2) How valid are the peer evaluations? How do the scores students receive from their peers compare to scores from an expert?
- (3) Regarding the construction of written explanations, how do students in classes using CPR assignments compare to students in similar classes who write explanations as part of regular homework without peer review?

III. METHODS

A. Study design

During Fall 2012 and Spring 2013, three of us [F. G., E. P., and S. R.] taught a total of 5 LEP classes ($N = 334$ students) at San Diego State University, California State University at San Marcos, and Tennessee Technological University. At two sites, LEP was a course for preservice elementary teachers; at the other site, it was a general education course taken by nonscience majors. Students in these courses were assigned five CPR tasks during the semester. We also administered a written explanation task on the final exam, and a multiple-choice, conceptual content knowledge assessment. As a comparison group, the written explanation and content knowledge assessment was also administered to students in 14 PET classes at five other colleges or universities; the results are described in Ref. [27]. Near the end of the Spring 2012 semester, semistructured interviews were conducted with six students (two each with high, middle, and low quiz grades) to gain insight into their experience with the CPR system.

For two of the five CPR tasks in the semester, we scored the students' texts using the same evaluation questions and rubric used by their peers (the evaluation questions are specific to each task). We refer to this researcher-generated score as a student's R score. Several undergraduate physics majors performed the scoring under our supervision. In the context of a conceptual course for nonscience majors, we

considered the physics majors to be experts. The following procedures were used. First, the scorers studied the prompt and evaluation questions, then scored the calibration responses. The results were compared and discussed until agreement was reached. Next, all scorers evaluated a common set of 10 student responses. Again, the results were compared and discussed. Finally, all scorers evaluated a second set of 10 student responses. The variation in these scores was taken as a measure of interrater reliability. After rater training, the interrater reliability was 90%; that is, the scorers agreed on 90% of the items scored. After this training period, the student work was divided randomly among the scorers, and one scorer rated each student's work. A similar training was conducted for each CPR task and a written explanation question given on the final exam. In general, the training was straightforward, and the rating process required little interpretation. This is in part a function of the nature of the task prompts and evaluation questions, which, as described below, were made very specific to facilitate the peer-review process.

With R scores for all students, we could then compare the score a student received from her peers to an expert's score. In particular, we compared the AWTR, which is the CPR system's best determination of student's text score, with R scores. This comparison included correlation testing of AWTR versus R score [15], a histogram of the differences between AWTRs and R scores, and equivalence testing [39,40]. The two one-sided test procedure was used for testing equivalence, with a significance value of $p = 0.05$. In this approach, we consider a confidence interval around the difference between the peer and expert scores. If this confidence interval lies within a predetermined equivalence margin, the peer and expert scores are equivalent at the $p = 0.05$ level. Since this method is essentially two one-sided tests, equivalence at the 0.05 significance level is established by considering a confidence interval of $(1 - 2 \times 0.05)100 = 90\%$ [39,40]. For this study, the equivalence margin was set, *a priori*, at ± 0.5 ; that is, based on pedagogical considerations, we decided that scores within 0.5 points (out of 10) were equivalent.

In addition to evaluating the validity of peer scoring on the CPR tasks, we investigated student performance on a written explanation question on the final exam, shown in Table I. The item asks for diagrams and a written narrative

TABLE I. The prompt for the written explanation task on the final exam.

Consider a bus that is leaving a bus stop. The driver takes his foot off the brakes and steps on the gas pedal to increase the engine power. The bus starts moving and speeding up. When it reaches the speed limit, the driver lets up on the gas pedal, but not completely, so that the engine power is reduced (but not to zero). The bus then continues moving at a constant speed.
Part 1. Draw two force diagrams: one for the time when the bus is speeding up, and one when it is moving at a constant speed. Note: Make sure that on each diagram you draw the appropriate speed arrows, paying attention to how their lengths compare. Also draw and label all relevant force arrows on the bus in both diagrams. Pay attention to the directions and relative lengths of the force arrows—both within each diagram, and between diagrams.
Part 2. Write an explanation for why the bus continues moving at a constant speed after the driver reduces the power.

TABLE II. The prompt and evaluation questions for the magnetism CPR task. In addition to the prompt, background material introduces the scenario and includes a video of a nail being magnetized by rubbing with a permanent magnet, and demagnetized by hammering. There were a total of 10 yes or no evaluation questions for the model of magnetism unit CPR task. Students are instructed to give an overall score equal to the number of yes responses to the evaluation questions.

Prompt for the magnetism task

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.]
-
-

The first 3 of 10 evaluation questions for the magnetism task

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 nail 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?
-
-

and was part of the final exam in the LEP classes. For comparison, the same item was given to students in PET classes. The question concerns Newton’s second law; LEP and PET each spend a unit on forces and motion, which in LEP includes a CPR task. For this study, the exam responses were scored using a 5 point rubric that was similar to the CPR evaluation questions and the same rater training procedures as the CPR tasks.

B. CPR tasks

The CPR tasks were developed, piloted, and revised before the semesters when this study took place. During Fall 2012 and Spring 2013, students were assigned five CPR tasks, an initial practice task, then one task each on interactions and energy, forces and motion, light, and a model of magnetism (each topic was a major unit in LEP). We estimate that it would take a conscientious student 1–3 hours to complete each of the four main tasks. The energy task required students to explain the energy transfers for a chain of interacting objects and determine the energy efficiency. The light task required students to use ray tracing to explain a prism system to correct a vision defect. The force task involved using Newton’s second law. Students were given the masses and some of the forces acting on a pair of carts, and asked to explain the results of a race between the carts. As part of the task, students were

asked to draw force diagrams, determine net forces, and calculate an unknown force using Newton’s second law. For the magnetism task, students read a description and watched a video of a magnetized nail being hit with a hammer so that it became unmagnetized. The prompt, shown in Table II, asks students to use their model of magnetism to explain the observed behavior. Table II also shows the evaluation questions for the magnetism CPR task. Students were instructed to score the response from 1 to 10 based on the number of evaluation questions to which they answered “yes.” In this study, we report on an analysis of the magnetism and force tasks. We selected these two tasks to analyze first because they represent the qualitative-quantitative range of CPR tasks in LEP. The magnetism task is more qualitative while the Newton’s second law task is more quantitative. Because the findings were consistent for both tasks, we did not analyze the other tasks assigned during the course. See the Supplemental Material for complete description of the magnetism task [41].

IV. RESULTS

A. Lessons learned

Based on pilot testing, we revised our initial versions of the CPR tasks, the way they are included in the course, and the way they are framed to students. These changes led to greater student success with the assignments (particularly

in the number of students successfully completing the calibrations), and fewer student complaints. During this process, activity theory served as a framework for the iterative process of understanding and refining our use of CPR, as discussed below.

A Fall 2011 pilot test led to a number of refinements in Spring 2012, including (i) more specific writing prompts and (ii) more focused or specific evaluation questions (see examples in Table II). The prompt includes specific directions on what is to be included (paragraph by paragraph) and even how the response should be formatted and presented. Having students' work in a standard form made the use of the evaluation questions much more straightforward. The evaluation questions are all yes or no questions that require minimal judgment. Because students may not know or be confident of an appropriate response to the prompt, each evaluation question makes clear what the correct answer is. The prompt and the evaluation questions are tightly coupled; each evaluation question asks students to check for the appropriate inclusion of something that the prompt directs the students to include.

In addition to revising the tasks themselves, we revised the way we explained and presented the CPR tasks to students. Because the tasks are complex and include many stages, a practice task was created so that students could get early experience with the process and structure of the CPR tasks, but with lower content demands and consequence for their grade. When describing the tasks to the students, we gave a general description of the procedures and the relevance to the course, but without the details of the procedures for each stage. Specifically, the CPR tasks were presented in the context of the curriculum's emphasis on science practices, the importance of constructing and evaluating explanations, and the role of peer review in science. This overview was provided in class at the beginning of the semester. Then, before each stage of the first CPR task, the class was given a more detailed description of the process for completing that stage. For subsequent CPR tasks, the instructor briefly reminded the students about the timing and instructions for each stage.

Activity theory draws our attention to the tools, roles, and rules in the activity, as well as their alignment and mediating relationships. The interaction between these elements, and the context in which they are embedded, can be complex. As a result, the way these elements are enacted in practice may differ from their formal or intended form. Tools are designed to be used in a particular way (with appropriate affordances and constraints), but they may be used very differently in practice. To effectively use the CPR system to support constructing and evaluating explanations, it is important that students use the system as intended, and take the calibration and review stages seriously. In activity theory terms, rules and roles must align with the tool and the context of its use. The changes described above increased this alignment. The result was a

better match between the students' level of expertise, the grading and weight of the assignment (a modest part of the whole course), and the time and effort required to complete the task. Similarly, explicitly emphasizing the importance of evaluation skills and peer review in science encouraged a classroom norm that supported students' role as an evaluator.

As a result of these changes, in the Spring 2012 implementation more students passed the calibrations, average calibration performance 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 their reaction to it), 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. Improvement remains possible, but taken together these results suggest that students better understood and more effectively used the overall CPR system, and their role as peer reviewer in particular.

B. Peer-score validity

To understand students' effectiveness in the role of peer reviewer, we compared their scoring to that of experts. If peer scoring is comparable to expert scoring, we expect both a small average difference between the peer and expert scores and a high correlation between the peer and expert scores. In this study, we evaluated the validity of the peer score by comparing the average weighted peer score and the R score using equivalence testing [39,40]. An equivalence margin of ± 0.5 was established, based on pedagogical considerations, so that scores within 0.5 points (out of 10) would be considered equivalent. Figure 3 shows histograms of the difference between the average weighted peer score and expert score on the magnetism and Newton's second law CPR tasks. A negative difference indicates that the expert score was higher than the average weighted peer score. For the magnetism task, the average difference was -0.08 ± 0.06 (average $\pm$ SEM). For the Newton's second law task, the average difference was -0.20 ± 0.06 (average $\pm$ SEM). The 90% confidence interval was -0.014 to 0.172 for the magnetism task and -0.300 to -0.104 for the Newton's second law task. For both tasks, the 90% confidence interval lies entirely within the range of the equivalence criteria, indicating that the average weighted peer score and the expert scores are equivalent at the $p = 0.05$ significance level [39,40].

FIG. 3. Histograms of the difference between the average weighted text score and expert score on the magnetism ($N = 313$, upper) and Newton's second law ($N = 312$, lower) CPR tasks.

We examined the correlation between peer score and R score for additional insight into their relationship. The results are shown in Fig. 4. Because a single expert scorer rates each submission on an integer scale, the R score is an integer, while the average weighted peer score is continuous. Based on a Pearson test, $R^2 = 0.82$ for the magnetism task, and $R^2 = 0.80$ for the Newton's second law task [42]. For both tasks, $P < 0.0001$, indicating a statistically significant correlation.

C. Impact on explanation performance

Besides the validity of the peer-evaluation process, we were interested in the impact of the CPR tasks on students' ability to construct scientific explanations. To assess this, an explanation question was included on the final exam in the LEP classes. The question concerned Newton's second law, and asked students to construct diagrams and a written narrative. As a comparison group, the same item was included on the final exam in PET classes. Homework and in-class activities in the PET curriculum include an emphasis on constructing and evaluating explanations of phenomena. However, the PET comparison classes did not use CPR tasks or any form of peer review; instead, students

FIG. 4. Research scores plotted versus average weighted peer scores for the magnetism ($N = 313$, upper) and Newton's second law ($N = 312$, lower) CPR tasks.

write explanations as part of their paper-and-pencil homework. LEP and PET each include a unit on forces and motion.

A 5-point rubric was used for scoring students' responses, following the same procedures used with the CPR tasks. Figure 5 shows average scores for the students in each curriculum, and Fig. 6 shows a histogram of scores.

FIG. 5. Average score on the final exam explanation item for LEP ($N = 334$) and PET ($N = 324$) students. Error bars are SEM.

FIG. 6. Histogram of post-test written explanation scores for LEP ($N = 334$) and PET ($N = 324$) students.

The difference between the averages is 1.2, which is statistically significant (based on a Mann-Whitney test, which was appropriate since the distributions were non-normal.). The difference between the average LEP and PET scores can be explained by the large number of students in the PET curriculum who received a score of zero. Of those zero scores, only 9.4% were students who did not provide a response. The remaining $\sim 90\%$ completed the explanation question, but did not receive any points on the scoring rubric.

The use of CPR tasks is a possible explanation for the difference in performance; a difference in student content knowledge is another. To examine the latter possibility, we assessed students' content learning gains in both courses. We administered a 12-item, multiple-choice conceptual assessment at the beginning and end of the semester. The items were representative of the material covered in both LEP and PET, and were developed by the project team in consultation with an external evaluator. Students took the pretest in class; the post-test was included as part of the final exam. The average gain (post-pre) for $N = 326$ matched LEP students was 4.2 items or 35%. The average normalized gain $[(\text{post} - \text{pre}) / (100 - \text{pre})]$ was 48.6%.

FIG. 7. Content assessment average percentage normalized gain for LEP ($N = 326$) and PET ($N = 309$) students ($p = 0.15$; $t = 1.456$ $df = 633$).

Based on a paired t test, the LEP students made statistically significant gains. A control group of PET classes also took the assessment under the same conditions. Figure 7 shows the average normalized gain for LEP and PET students. In both curricula, the average normalized gain is in the mid to upper 40%, suggesting that students in both courses learned nearly half of what they could have. The difference between LEP and PET students is not statistically significant.

V. DISCUSSION

Our experience indicates that CPR tasks can be used in a conceptual physics course to engage students in the science practices of constructing and evaluating explanations of physical phenomena. This requires students to take on a different role, that of evaluator. This is challenging because students' understanding of the material is still developing. This is addressed by the calibration process within CPR, as well as the design of the tasks. Successful implementation required prompts and evaluation questions that were highly structured and specific. Furthermore, students generally do not expect to take on, nor have experience with, the role of evaluator in a classroom context. To facilitate students' shift into this new role, we found it important to frame and motivate the tasks in the context of the curricular goals, as well as provide frequent and detailed guidance to students on how to complete the CPR tasks. After improving the alignment between the roles, norms, and the use of the CPR tool, we encountered very few problems with incorporating the CPR tasks in the curriculum, and the peer grading process via the CPR system required little instructor involvement. Furthermore, for the two CPR tasks analyzed in this study, the average weighted scores students received from their peers were highly correlated with expert scores, and the average weighted peer scores were statistically equivalent to expert scores. We conclude that students are able to effectively take on the role of evaluator. Further, even in large classes, CPR tasks are an effective and practical way to engage students in the practices of constructing and evaluating explanations. However, our activity theoretic lens draws attention to the complexity of the CPR tool and the sensitivity of the system to context. We caution that these results may not generalize; the validity of the peer scoring process in CPR will depend on the student population, the curriculum, and especially the task design.

The standard for constructing and evaluating explanations in the CPR tasks included in this study arguably constitute a minimum expectation. A higher standard would require that students effectively respond to more open prompts. This would require greater sophistication from students and provide them with more flexibility to express their own thinking and ground their response in their own experiences. Similarly, tasks could include evaluation questions that require greater judgement and interpretation. Given the nature of the tasks, however,

students' ability to perform this kind of evaluation is tightly linked to their content knowledge, which is inexpert and varies from student to student. An important feature of the CPR tasks is the degree to which each student is affected by other students in the course. If effectively responding to the CPR prompt and applying the evaluation questions requires a high degree of sophistication (relative to the class average), lower performing students will be less likely to perform effective evaluations of their peers. In this way, CPR tasks are different from traditional assignments where each student performs on his or her own. This change in the interactions between students (characterized as a change in roles and division of labor in activity theory) is a significant stress and potential point of breakdown. If students are unwilling or unable to effectively evaluate, feel this is an inappropriate role, or simply lack confidence in their peers, then peer review is unlikely to be successful. As instructors and researchers, we must be sensitive to the challenge posed by changing these roles. Highly structured prompts and specific evaluation questions in the CPR tasks help provide the support needed for this change.

One might expect that with practice students would improve at constructing and evaluating explanations, and therefore the CPR tasks could increase in sophistication over time during the course. In this study, all the CPR tasks had a similar degree of structure and specificity, so we do not have evidence to address this question. We designed the tasks this way in part because each CPR task focuses on a different topic. As noted above, performance on the CPR tasks requires specific content knowledge; constructing and evaluating explanation are not abstract skills, but are grounded in the context of the subject matter. This is consistent with the NGSS's integration of practices, disciplinary ideas, and crosscutting concepts: "we use the term 'practices,' instead of a term such as 'skills,' to stress that engaging in scientific inquiry requires coordination both of knowledge and skill simultaneously [12]." As a result, assessing growth in students' ability to construct and evaluate explanations is a complex task that requires further investigation.

We have some evidence that including explanation and evaluation in the LEP curriculum through CPR tasks results in better performance on explanation items, compared with including explanation and evaluation in the PET curriculum through traditional assignments. Strikingly, a large portion (~35%) of PET students scored zero on a final exam item requiring an explanation. Yet, conceptual learning gains were similar in both courses. Within an AT framework, we explain these results by comparing the tools, rules and norms, and roles that mediate students' engagement with the explanation tasks during the course. CPR tasks provide greater degrees of structure and feedback compared to traditional assignments. PET includes pencil and paper homework assignments, which include questions requiring students to explain phenomena. Different instructors may

have different grading practices, and provide varying degrees of feedback to students. Even when feedback is provided, students may not study this feedback. The structure of CPR tasks provides multiple opportunities for students to receive and reflect on feedback; furthermore, their performance on subsequent tasks (evaluating other students' work) is contingent on their ability to incorporate this feedback. This suggests that the calibration and review stages play an important role in students' learning from the CPR tasks. While the specific contribution of the different stages is a topic for further investigation, we interpret this in terms of the functionality and affordances of CPR as a mediating tool that restructures the assignment, changes roles, and enables activities that would be difficult using paper and pencil assignments.

Several important factors should be considered when interpreting the LEP-PET comparison, however. First, while the LEP and PET curricula cover very similar topics and were designed using the same research-based pedagogical principles, the format of two curricula are different [27,30]. Second, the frequency and intensity of explanation tasks is also different. In PET, brief explanation tasks are regularly included in class activities and homework assignments. In LEP, students were regularly asked to evaluate sample explanations in class, but constructing explanations only took place during the CPR tasks, which were assigned approximately every 3 weeks and took about 1.5 weeks to complete (as students worked through the sequential text entry, calibration, and review stages). A more direct evaluation of the impact of the CPR tasks, and a determination of the contribution of the different stages (writing, calibration, peer review) is needed.

VI. CONCLUSIONS

We have used CPR tasks in a conceptual physics course to engage students in the practices of constructing and evaluating explanations. Using tasks with specific, structured prompts and evaluation questions, the peer scoring process in CPR is equivalent to expert scoring. Furthermore, on a written explanation item on the final exam, students in courses using CPR outperformed students in similar courses using traditional writing assignments without a peer evaluation element. We therefore regard CPR as a valuable and effective way to include the science practices of constructing and evaluating explanations, even in large classes.

ACKNOWLEDGMENTS

The authors would like to acknowledge Melissa Dancy, CSUSM Instructional and Information Technology Services, Paul Heft, Chase Keene, Kornel Czarnocki, Madison Perry, Nate Reynolds, and Heath Kirkwood. This work was supported by NSF DUE Grants No. 0717791 and No. 1044172.

- [1] This is an instance of a more general need for scalable instructional models that can allow meaningful engagement and feedback in large classes. Peer Instruction and SCALE-UP are well-known examples.
- [2] C. Piech, J. Huang, Z. Chen, C. Do, A. Ng, and D. Koller, in *Educational Data Mining 2013* (International Educational Data Mining Society, 2013), <http://www.educationaldatamining.org/IEDMS/EDM201>.
- [3] S. P. Balfour, Assessing writing in MOOCs: Automated essay scoring and Calibrated Peer Review, *Res. Prac. Assessment* **8**, 40 (2013).
- [4] S.-Y. Lin, S. Douglas, J. Aiken, C.-L. Liu, E. F. Greco, B. Thoms, M. Caballero, and M. Schatz, *Peer Evaluation of Video Lab Reports in an Introductory Physics MOOC*, 2014 PERC Proceedings, Minneapolis, MN (AAPT, College Park, MD, 2014), pp. 163–166.
- [5] S. Douglas, S.-Y. Lin, J. Aiken, B. Thoms, E. F. Greco, M. Caballero, and M. Schatz, *Peer Evaluation of Video Lab Reports in a Blended Introductory Physics Course*, 2014 PERC Proceedings, Minneapolis, MN (AAPT, College Park, MD, 2014), pp. 75–78.
- [6] Calibrated Peer Review was developed at UCLA. More information is at <http://cpr.molsci.ucla.edu>.
- [7] E. F. Redish, J. M. Saul, and R. N. Steinberg, Student expectations in introductory physics, *Am. J. Phys.* **66**, 212 (1998).
- [8] E. Etkina, A. Van Heuvelen, S. White-Brahmia, D. T. Brookes, M. Gentile, S. Murthy, D. Rosengrant, and A. Warren, Scientific abilities and their assessment, *Phys. Rev. ST Phys. Educ. Res.* **2**, 020103 (2006).
- [9] J. L. Lemke, *Talking Science: Language, Learning, and Values* (Ablex Publishing Corp., Norwood, NJ, 1990).
- [10] R. Driver, P. Newton, and J. Osborne, Establishing the norms of scientific argumentation in classrooms, *Sci. Educ.* **84**, 287 (2000).
- [11] F. J. Rutherford and A. Ahlgren, *Science for all Americans* (Oxford University Press, Oxford, United Kingdom, 1991).
- [12] National Research Council, *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas* (National Academies Press, Washington, DC, 2011).
- [13] NGSS Lead States, *Next Generation Science Standards* (National Academies Press, Washington, DC, 2013).
- [14] N. Falchikov and J. Goldfinch, Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks, *Rev. Educ. Res.* **70**, 287 (2000).
- [15] P. M. Sadler and E. Good, The impact of self-and peer-grading on student learning, *Educ. Assessment* **11**, 1 (2006).
- [16] M. van Zundert, D. Sluijsmans, and J. van Merriënboer, Effective peer assessment processes: Research findings and future directions, *Learning Instr.* **20**, 270 (2010).
- [17] S. Freeman and J. W. Parks, How Accurate Is Peer Grading?, *CBE Life Sci. Educ.* **9**, 482 (2010).
- [18] J. R. Harris, Peer assessment in large undergraduate classes: an evaluation of a procedure for marking laboratory reports and a review of related practices, *Adv. Physiol. Educ.* **35**, 178 (2011).
- [19] D. E. Paré and S. Joordens, Peering into large lectures: examining peer and expert mark agreement using peerScholar, an online peer assessment tool, *J. Comput. Assist. Learn.* **24**, 526 (2008).
- [20] K. Topping, Peer assessment between students in colleges and universities, *Rev. Educ. Res.* **68**, 249 (1998).
- [21] K. Cho, C. D. Schunn, and R. W. Wilson, Validity and reliability of scaffolded peer assessment of writing from instructor and student perspectives, *J. Educ. Psychol.* **98**, 891 (2006).
- [22] Y. H. Cho and K. Cho, Peer reviewers learn from giving comments, *Instr. Sci.* **39**, 629 (2011).
- [23] M. E. Walvoord, M. H. Hoefnagels, D. D. Gaffin, M. M. Chumchal, and D. A. Long, An analysis of Calibrated Peer Review (CPR) in a science lecture classroom, *J. Coll. Sci. Teach.* **37**, 66 (2008).
- [24] L. D. Margerum, M. Gulsrud, R. Manlapez, R. Rebong, and A. Love, Application of Calibrated Peer Review (CPR) writing assignments to enhance experiments with an environmental chemistry focus, *J. Chem. Educ.* **84**, 292 (2007).
- [25] L. Likkel, Calibrated Peer Review essays increase student confidence in assessing their own writing, *J. Coll. Sci. Teach.* **41**, 42 (2012).
- [26] A. A. Russell, Calibrated Peer Review: a writing and critical-thinking instructional tool, *Invention and Impact: Building Excellence in Undergraduate Science, Technology, Engineering and Mathematics (STEM) Education* (American Association for the Advancement of Science, Washington, D.C., 2004).
- [27] E. Price, F. Goldberg, S. Robinson, D. Harlow, M. McKean, S. Keene, and K. Czarnocki, *Development and Evaluation of a Large-Enrollment, Active-Learning Physics Curriculum*, 2013 PERC Proceedings, Portland, OR (AAPT, College Park, MD, 2013), pp. 285–288.
- [28] F. Goldberg, E. Price, S. Robinson, D. Boyd-Harlow, and M. McKean, Developing the learning physical science curriculum: Adapting a small enrollment, laboratory and discussion based physical science course for large enrollments, *Phys. Rev. ST Phys. Educ. Res.* **8**, 010121 (2012).
- [29] F. Goldberg, S. Robinson, and V. Otero, *Physics and Everyday Thinking* (It's About Time, Mount Kisco, NY, 2005).
- [30] F. Goldberg, V. Otero, and S. Robinson, Design principles for effective physics instruction: A case from physics and everyday thinking, *Am. J. Phys.* **78**, 1265 (2010).
- [31] F. Goldberg, S. Robinson, R. Kruse, N. Thompson, and V. Otero, *Physical Science and Everyday Thinking* (It's About Time, Mount Kisco, NY, 2008).
- [32] F. Goldberg, S. Robinson, E. Price, D. Harlow, and M. McKean, *Learning Physical Science* (It's About Time, Mount Kisco, NY, 2012).
- [33] This version is called *Next Generation Physics and Everyday Thinking*. The materials are described at <http://petcyberpd.iat.com>.
- [34] Y. Engeström, *Learning by Expanding: An Activity-Theoretical Approach to Developmental Research* (Orienta-Konsultit, Helsinki, 1987).
- [35] M. Cole, *Cultural Psychology: A Once and Future Discipline* (Harvard University Press, Cambridge, MA, 1998).
- [36] B. A. Nardi, *Context and Consciousness: Activity Theory and Human-Computer Interaction* (MIT Press, Cambridge, MA, 1996).

- [37] V. Kaptelinin and B. A. Nardi, *Acting with Technology: Activity Theory and Interaction Design* (MIT Press, Cambridge, MA, 2009).
- [38] D. Norman, *The Psychology of Everyday Things* (Basic Books, New York, 1988).
- [39] S. Wellek, *Testing Statistical Hypotheses of Equivalence and Noninferiority* (CRC Press, Boca Raton, 2010).
- [40] E. Walker and A. S. Nowacki, Understanding equivalence and noninferiority testing, *J. Gen. Internal Medicine* **26**, 192 (2011).
- [41] See Supplemental Material at <http://link.aps.org/supplemental/10.1103/PhysRevPhysEducRes.12.020145> for full description of the magnetism task.
- [42] A nonparametric Spearman correlation test also indicates a statistically significant correlation.