Improving performance in upper-division electricity and magnetism with explicit incentives to correct mistakes

Andrew J. Mason,[1] Jessica M. McCardell,[2,3] Philip A. White,[4,5] and John S. Colton[2]

[1]*Department of Physics and Astronomy, University of Central Arkansas, Conway, Arkansas 72035, USA*
[2]*Department of Physics and Astronomy, Brigham Young University, Provo, Utah 84602, USA*
[3]*Department of Curriculum and Instruction, University of Idaho Coeur d'Alene, Coeur d'Alene, Idaho 83814, USA*
[4]*Department of Statistics, Brigham Young University, Provo, Utah 84602, USA*
[5]*Berry Consultants, Austin, Texas 78746, USA*

(Received 3 January 2023; accepted 15 June 2023; published 18 July 2023)

This study seeks to determine whether giving an explicit incentive to students in an upper-division first-semester electromagnetism course (EM1), in the form of partial credit for reworking unit exam problems, will improve their problem-solving skills as measured by performance on identical problems on the final exam. Three problems–a primarily algorithmic problem, a primarily conceptual problem, and a problem that blended conceptual with algorithmic–were selected and analyzed over the course of three consecutive sections of EM1, for which each student could freely choose whether or not to rework mistakes on each one of the given problems in exchange for the aforementioned partial credit. Regression models were chosen for quantitative analysis, with the covariate being unit exam performance and whether the student reworked the problem. Results indicate that overall, students who choose to rework problems perform better on the final exam attempt, at a level that often does not correlate strongly with unit exam performance, whereas students who decline to rework problems have a stronger correlation between unit exam and final exam performances. The results show a clear difference between the two stages of problem-solving, namely, invoking the correct principles and applying the principles, where the latter showed a more significant effect with a much larger effect size. Qualitative analysis of a sample of students interviewed about their exam solutions showed that reworking an exam problem for some students did result in more expertlike problem-solving trends; that being said, there were instances of persistent novicelike trends regardless of reworking, as well as instances of students independently reviewing unit exam problems in preparation for the final even though the partial credit incentive was declined at the time of the unit exam. Thus, while the intervention showed overall benefits insofar as exam performance is concerned, the usage of more well-defined scaffolding, e.g., tutorials, may prove a more thorough and definite benefit for improving problem-solving skills.

DOI: 10.1103/PhysRevPhysEducRes.19.020104

I. INTRODUCTION

A wealth of literature exists regarding innovations for improving student problem-solving skills in physics that rely on topics of cognitive psychology [1,2]. Background studies at the introductory level have expanded the topic of problem-solving into specific subtopics within problem-solving, e.g., metacognition on problem-solving reflection [3,4] and drawing diagrams [5]; related areas of physics education research, e.g., epistemic attitudes toward problem-solving [6] and transfer of learning for similar principles between problems [7]; and innovations to improve problem-solving at the introductory level, e.g., reflection on problem-solving attempts [3,8], recommended group dynamics [9], and problem types such as context-rich problems [10,11] and open-ended/ill-defined problems [11].

Problem-solving investigations at the upper-division level of physics have tended to be relatively less frequent than at the introductory level. There are investigations into specific topics within fields such as quantum mechanics [12] and thermodynamics [13]; preliminary measurements of topics such as metacognition [14] and attitudes toward upper-division problem-solving [15], as well as investigations aimed at improving students' experiences in said topics [16]; and tutorials aimed at understanding concepts behind specific topics, which may lend themselves well to problem-solving [17]. In addition, researchers have noticed the internal mental conflicts that advanced physics students

may have between how they perceive aspects of physics learning and how they carry out those aspects. This can affect how they learn advanced physics themselves [14], and even how, as teaching assistants, they can differ between their views of grading as formative assessment and their actual practices of grading as summative assessment [18].

Brown *et al.*'s initial success [16] at helping students improve their understanding of quantum mechanics (QM) problems inspire this paper's creation, as the authors seek to replicate the Brown *et al.*'s study for upper-division electricity and magnetism (EM) problems and expand understanding of the effect through additional statistical analysis and student interviews. A preliminary study of quantitative data from having students rework problems in upper-division EM [19] suggested that students who take advantage of incentivization, namely partial credit restored on corrected problems, to rework mistakes on unit exam problems will perform better on the same problems given verbatim without prior notice on the final exam. The authors seek to expand upon this study with the following aims to establish more clearly whether there is a real quantitative benefit to reworking the problems, as opposed to a priming effect due to further experience with the specific problems; and to establish from qualitative think-aloud protocol interview data of student volunteers whether there are specific reasons behind this effect.

A. Common concerns for upper-division EM and QM

As also discussed by Brown *et al.* [16], many students do not take advantage of problem-solving as a learning opportunity at the upper-division level of physics [20]. This can cause a performance gap between students who will voluntarily learn from their problem-solving attempts and students who will not; addressing this issue properly therefore will improve overall pedagogical efforts [21]. Prior research shows the effectiveness of rewarding students for self-monitoring and correcting their mistakes [3,20]. There is also a recognition that assumptions are frequently made about upper-division physics students' ability to already have well-developed problem-solving skills; while a few studies, e.g., Mason and Singh [14], show that these assumptions are unsubstantiated, there is a need for more precise investigations as to what precisely upper-division students' problem-solving habits may entail. There is a relatively low volume of literature on researching problem-solving skills of upper-division skills; among the literature that exists on this topic, much focus is given to specific topics, e.g., QM [12,14,15,22–24], or to specific issues with problem-solving in upper-division courses, e.g., mathematics [25] and computation [26]. With regard to literature specifically on problem-solving in EM, there are items with regard to investigating conceptual understanding [17] and mathematical reasoning [25] in upper-division EM. Part of this paper's intention is to add to the findings of Brown *et al.*, as well as to the literature on problem-solving in upper-division EM, by attempting to examine explicit problem-solving with a mixed methods approach, specifically a quantitative approach with follow-up qualitative interviews.

A common concern regarding what constitutes a "problem-solving framework" at the upper-division level. Problem-solving frameworks [27] are routinely discussed in relation to introductory physics courses, but not frequently at the upper-division level. Concerns such as preparing students more explicitly for graduate school and/or the workforce outside of academia with open-ended problems [2] becomes a concern as well for upper-division instructors.

B. Additional concerns specific to electricity and magnetism

One point of interest that may cause upper-division EM results to differ from that of QM is the relative nature of the two subjects. For example, certain specific topics are covered in multiple upper-division courses, e.g., equipotential lines in both electrostatics and thermodynamics [28]. Student understanding of such topics can be affected by the respective contexts of both courses' overall content. QM often must present new concepts to students as well as new mathematical methods and other problem-solving skills. Students who are seeing quantum concepts for the first time may struggle to accommodate the learning of these new concepts alongside applying mathematical methods that can also be novel (e.g., Dirac notation) [29]. On the other hand, EM is typically taught in an introductory second-semester course and then concepts from that course are reintroduced with additional math at the upper-division level. This difference between subjects may translate to changes in cognitive load for upper-division students [30]. The degree to which this is similarly an issue for upper-division electromagnetism needs to be explored; students typically have a foundational layer of conceptual understanding at the introductory level, to the point where mathematical methods may prove to be overemphasized at the upper-division level. Whether the difference in domain-specific content knowledge between EM and QM will cause differences in effect from an otherwise similar intervention remains unclear. In other words, one needs to see how the Brown *et al.* methodology for QM does and does not work similarly for EM.

Another point of concern is that the ability to treat problem-solving in upper-division EM with well-validated research products remains relatively unsupported by the PER community. While conceptual tutorials exist from multiple sources for upper-division electromagnetism [17], as mentioned above, they are primarily focused on conceptual understanding; an explicit application toward problem-solving skills remains relatively sparse and focused upon specific topics [31]. Therefore, a longer-term goal beyond this project would be to organize currently recognized methods to explore electromagnetism

problem-solving skills—e.g., applying well-understood aspects of cognition: worked examples [32], metacognition in problem-solving [3], and reinforcing individual aspects of a problem-solving framework [5,27]—into designed materials for EM that are more explicitly focused on improving problem-solving, similar to what has been done for QM [29].

C. Research questions from the authors

In this paper, we use the Brown *et al.* [16] model of allowing students a chance to rework their solutions on unit exam problems, in exchange for course credit as an incentive, in order to allow for a group comparison between students who choose to rework each solution for which they are eligible to gain a substantial amount of exam grade, and students who choose to refrain from reworking solutions for credit even if they are similarly eligible. Our research questions are as follows:

First, we will quantitatively establish whether incentivized reworking of unit exam problems shows improvement, with both statistical significance and a meaningful effect size, for eligible students who accept the opportunity, as opposed to eligible students who decline it. This is measured by choosing three unit exam problems for which students generally struggled in particular, repeating them verbatim on the final exam without advance notice, and examining pre-post gains for those three specific problems accordingly. A linear regression model will be used to examine trends of final exam performance on each problem vs respective unit exam performance. The confounding variable of students' unit exam scores, as an indication of how much help they needed per problem, will be addressed as a covariate.

Second, the authors seek to more closely address the effect of conceptual vs algorithmic problem-solving, namely whether the nature of the problem in question has an effect on invoking correct principles and on correctly applying the principles to obtain a solution. The three problems in this study include a mixture: one is primarily conceptual, one is primarily algorithmic, and one is a mixture of both.

Third, the authors seek to add insight into the first two questions via qualitative interviews of a randomly selected subset of the student sample. The interviews consist of using a think-aloud protocol in which students discuss their own unit exam attempts and final exam attempts for each problem in the study, as available. The authors' third research question is whether patterns of responses emerge for each of the three problems that may help to explain the quantitative results shown for the previous research questions.

II. METHODOLOGY

A. Student population

The host university for the study was a large private research university in the Mountain West region of the United States. The first-semester course of the upper-division electromagnetism sequence at this university, electricity and magnetism 1 (EM1), was selected for analysis for the study. Three sections for this course were selected for analysis: the Fall 2019 section (hereafter referred to as "F19"), the Spring 2021 section (S21), and the Fall 2021 section (F21). The S21 section was in fact the first of two 7.5-week summer terms, in which instruction proceeded at twice the rate of the regular 15-week full semester terms for F19 and F21. The timing of the semesters avoided the Covid-19 protocols as full in-person instruction was permitted.

In each section, students were provided with a consent form to sign, which granted permission for data collection and analysis in the study; students who declined to sign were omitted. In the F19 semester, 25 students (out of 27 overall) consented to participate, as did 14 students (out of 19 overall) in the S21 semester and 36 students (out of 44 overall) for the F21 semester, for a total of 75 participants across all three sections.

B. Course structure

For each course, the instructor used a traditional lecture format (50-minute lectures, three lectures per week) with specific active learning components added. For example, formative assessments [33] at the beginning of each class, in the form of conceptual quizzes that were not for credit, began each class in order to review previous topics and introduce new topics for the given lecture. Socratic dialogue was used during the lecture periodically in questioning called-upon students. Students were sometimes given assignments to do in class either individually or in pairs, e.g., student volunteers to assist the class with worked examples. The textbook used for the course was by Griffiths [34], and topics covered included vector analysis; electrostatics; special techniques for calculating potentials; electrostatic fields in matter; magnetostatics; magnetostatic fields in matter; Faraday's Law; and Maxwell's correction to Ampere's Law.

The three unit exams for the course were given at a dedicated testing center outside of regular class time. Students were permitted to take the exam at any time during a scheduled number of days: five days for the regular-semester courses (F19 and F21) and three days for the summer term (S21). Once students began the exam there was no specific time limit to finish, apart from the close-of-day of the testing center, but the students were required to complete the exam in a single sitting. The cumulative final exams were either given at a specified place and time with a soft three-hour time limit (F19 and F21) or again at the testing center during a two-day window under similar conditions as the unit exams (S21). In all cases, students were allowed to bring a calculator and a single two-sided page of handwritten notes and formulas to each exam.

C. Procedure

Per Brown *et al.* approach [16], after students received their graded exams for each unit exam, they were permitted an opportunity to rework their mistakes on any of the exam problems they wished and resubmit the corrected work in exchange for partial credit (50% of any recovered points). The deadline for resubmission was typically within five days of the exams being initially returned to the students with their grades. Students could choose to rework some problems but not others, as desired.

A key difference from Brown *et al.* approach is that the experimental group and control group for each problem was defined by whether a student chose to rework a given problem (experimental) or declined to rework it (control). The Brown *et al.* approach in contrast had separate course sections serving, respectively, as experimental and control groups. There are two reasons for the departure from this aspect of the Brown *et al.* model. First, the authors initially were unsure how many sections of EM1 would be available for the study and had to adapt to the possibility that only one section would be available within a reasonable time frame. Upon being able to secure other sections as the study continued, this concern was of course obviated; however, the authors chose to retain this aspect of the study to maintain consistency as much as possible across all sampled sections.

Second, due to the main result in Brown *et al.* that the experimental group seemed to benefit from reworking problems relative to the control group, the decision was made that the better ethical choice was to give all students involved in the study the option to rework or not rework. As a result, in this study, students were not necessarily uniformly part of either group across all sampled problems, as was the case in Brown's study.

Third, Brown *et al.* mention that "All students who had less than perfect score on the midterm exams took advantage of the incentive to correct their mistakes for course credit." [16] (p. 4). This was not the case for the student populations in this study, in which students declined to rework mistakes for the unit exam attempt on each of the chosen problems on 50 out of 99 total occasions (in which we define an "occasion" as one student deciding for one problem; see Table I), making for groups where there were 49 occasions in which eligible students would rework a problem. Therefore, Brown *et al.* comparison is between students who had the option to rework exam problems and chose that option uniformly when possible (i.e., when not scoring 100% on a given unit exam problem), and students who did not have the option to rework exam problems due to being in the control group sections, as the student population contained no subset of students in the incentivization group who declined the opportunity to rework when it was available. Accordingly, Brown *et al.* could interpret their students in terms of average gain for each student across all four problems used in each of the four sections.

In contrast, the incentivization structure for this study required sorting students into three groups on a per-problem basis: having the opportunity to rework mistakes on a given problem and choosing to do so (the control group); having the opportunity to rework mistakes on said problem, but declining to do so (the experimental group); and not having much opportunity to rework mistakes due to scoring 90% or better on the unit exam version of said problem (omitted from analysis). For example, a given individual student could be identified as part of the experimental group for P1 by choosing to rework it and then identified as part of the control group for P2 by choosing not to rework it.

For the cumulative final exam, three prior unit exam problems were chosen to be featured again verbatim without informing the students. For the S21 and F21 semesters, the three problems chosen are featured in Fig. 1, respectively, labeled P1, P2, and P3. The three problems have different qualitative characteristics as follows: P2 is primarily a conceptual problem, requiring students to think carefully about the behavior of electric field lines in order to first sketch them in a two-dimensional space around two unequal point charges, then graph the horizontal component (E_x) as a function of horizontal position (x). P3 on the other hand is primarily an algorithmic problem, in which students must recognize and calculate each of three multipolar components of the electric potential at a specific spot in three-dimensional space near a given charge distribution. P1 has a primarily algorithmic part, namely calculating the change in electric potential by changing the distance between two charges in a given dipole, and a primarily conceptual part, namely explaining in words whether that change in electric potential is positive or negative. For the F19 semester, P1 was used, along with two other problems that ended up being discarded due to a ceiling effect that precluded substantial improvement from the unit exam to the final exam.

D. Rubrics for problem-solving attempts

Rubrics were constructed in order to assess each of the three problems featured in the study on both the unit exam attempts and the final exam attempts. One author designed rubrics for each of the three featured problems. As an example, Fig. 2 shows the rubric constructed for P1; an example is shown with scores from a hypothetical student. Each problem's rubric was constructed to identify each item that a student would have to recognize and correctly express in order to receive full credit for each portion of a problem. The items were placed into one of two categories: invoking the correct concepts (labeled as "invoked" items) and applying the concepts correctly (labeled as "applied" items). For P1 in Fig. 2, for example, the two concepts that

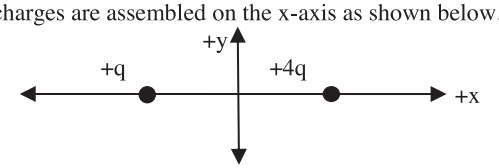
Name	Unit exam	Text
P1	Exam 1 (F19, S21, F21)	Two point charges of equal but opposite charge are separated by a distance d, the +q charge being on the left and -q on the right. If the charges are each moved a distance d/2 away from each other, what is the change in potential energy of the system? Specify whether the potential energy has increased or decreased, and give a conceptual explanation for why this is the case.
P2	Exam 1 (S21, F21)	Two unequal charges are assembled on the x-axis as shown below.  a) Make a sketch of the electric field lines on the figure. b) In the space below make a rough plot of E_x vs. x for points along the x-axis. Don't worry about any numbers, just the general shape of what happens to the x-component of the field – which can be positive or negative depending on whether E points to the right or the left.
P3	Exam 2 (S21, F21)	A uniform charged rod with linear charge density +λ lies on the z-axis, extending from z = 0 to z = –d as shown. Suppose you want to calculate the potential at point P which lies on the positive z-axis. a) What is the monopole contribution to the potential at point P, V_{mono}? b) In order to get a little more accuracy than the monopole potential, consider the dipole potential: $V \approx V_{mono} + V_{dip}$. (You may or may not have noticed, but the dipole formula does not actually require there to be both positive and negative charges, although typically we think of dipoles in those terms.) What is the dipole contribution to the potential at point P, V_{dip}? c) In order to get even a little more accuracy than the monopole and dipole potentials combined, consider the quadrupole potential: $V \approx V_{mono} + V_{dip} + V_{quad}$. What is the quadrupole contribution to the potential at point P, V_{quad}?

FIG. 1. The three problems chosen for analysis in this study, as well as which unit exam (of three, not including final) each problem was featured in. Included in parentheses is the semester(s) for which each problem was used.

needed to be invoked were, respectively, the potential energy between two objects and how that potential energy could change. For applying, the three applications that were needed were to apply potential energy before and after a change had occurred (namely the change in separation of charges); to identify the changed distances correctly (as students frequently made mistakes on this particular application); and to perform the algebra correctly after having set up the approach to it. By way of comparison to P1 as shown, P2 had nine invoked items and two applied items,

General criteria	**Specific criteria**	**Sample Score**
Invoked Concepts	Potential energy between two objects	1
	Conceptual understanding of increase/decrease of PE	0.5
Invoked Score (out of 6 pts)		1.5/2 = 0.75 = 75.0% 0.75 × 6 = ***4.50***
Applied Concepts	PE formula applied to "before separation" and "after separation" situations	0.67 (rounded from ⅔)
	Correctly identifying distances from the given information	1
	Correctly doing algebra to arrive at answer	1
Applied Score (out of 6 pts)		2.67/3 = 0.890 = 89.0% 0.89 × 6 = ***5.34***
Total Score (out of 12 pts)		***4.50*** + ***5.34*** = ***9.84*** 9.84/12 = 0.82 = **82.0%**

FIG. 2. Sample rubric for P1 for hypothetical student; see Fig. 1 for P1 problem statement.

TABLE I. Summary of students who, respectively, were and were not eligible for reworking problems on each featured problem in the unit exams and final exams. See Fig. 1 for the text associated with each problem ID number.

		Eligible? (< 90% on unit exam attempt)			
Semester (Total No.)	Problem	Yes, reworked	Yes, did not rework	No	Total eligible attempts (reworked + no rework)
F19 (25)	P1	8	7	10	15
S21 (14)	P1	4	4	6	8
	P2	7	2	5	9
	P3	4	1	9	5
F21 (36)	P1	6	17	13	23
	P2	11	6	19	17
	P3	9	13	14	22
All	All	49	50	76	99

and P3 had six invoked items and seven applied items, due to the different structure and added complexity for those problems.

To rate students' attempts, each rubric item was marked on a 5-point scale from 0 to 1 as follows: 0 = completely wrong, 0.33 = more wrong than right but not completely wrong, 0.5 = partially right and partially wrong, 0.67 = more right than wrong but not completely right, and 1 = completely right. The overall invoked score for each student averaged all invoked items' scores and multiplied by a normalized value (in P1's case for example, out of 6 points), and a similar process was done with all applied items' scores to determine the overall applied score (also out of 6 points for P1). The overall score was the sum of the normalized invoked score and normalized applied score (in P1's case, out of $6 + 6 = 12$ total points).

Another author and external third parties (see Acknowledgments) were consulted in order to establish a mutually agreed validity for these rubrics. To ensure interrater reliability, two raters graded each student independently for a sample of students on P1 and two additional problems (not used), then compared ratings. For the problems compared, 81.5% initial agreement was achieved prior to any discussion between raters, and ultimately discussion and double-checking evaluations improved agreement to 90.8%. Upon establishing interrater reliability, rubrics could then be used for all students on both unit exam attempts and final exam attempts.

E. Statistical methodology

To analyze the quantitative data, for each problem, students were categorized as being in one of three groups: those who chose to rework that problem after a unit exam attempt ("did rework"), those who declined the opportunity to rework for partial credit ("did not rework"), and those who had performed 90% or better on the unit exam attempt and were excluded from consideration for this analysis regardless of whether they had reworked or not on the basis that they had already mastered the material. Table I shows the populations of the students in the three groups for each of the three problems across each of the three sections of the course. Note that a given student could potentially be in the did not rework group for one problem and in the did rework group for another problem.

To answer the first research question of the overall benefit of reworking, eligible students who chose to rework a given problem are compared to eligible students who declined to rework the problem. To answer the second research question, each problem's traits of being algorithmic or conceptual are considered separately in order to determine benefits accordingly: P1 being a mixture of conceptual and algorithmic, P2 being primarily conceptual, and P3 being primarily algorithmic.

The possible differences in performance of the did rework and did not rework groups are assessed using multiple regression models for predicting the final exam score, based on unit exam scores and whether or not students reworked the problem as explanatory variables. When aggregating the three problems together, because a student may rework more than one problem we additionally use mixed effects models that have fixed effects for the students' unit exam scores and whether they reworked the problem, as well as random effects for each student. The random effects for student account for the correlation between scores for the same student. Both model types are fit using maximum likelihood estimation.

We use analysis of covariance (ANCOVA) to compare three regression models via two tests, which we call Tests A and B. We use ANCOVA to test the effect of reworking, as well as the interaction between reworking and unit exam scores while accounting for possible differences in unit exam scores between the two groups. The simplest regression model uses only the unit exam scores as the predictor of final exam scores, and we call this the "exam only" model. We do not present the results for the "exam only" model because it does not address our research question.

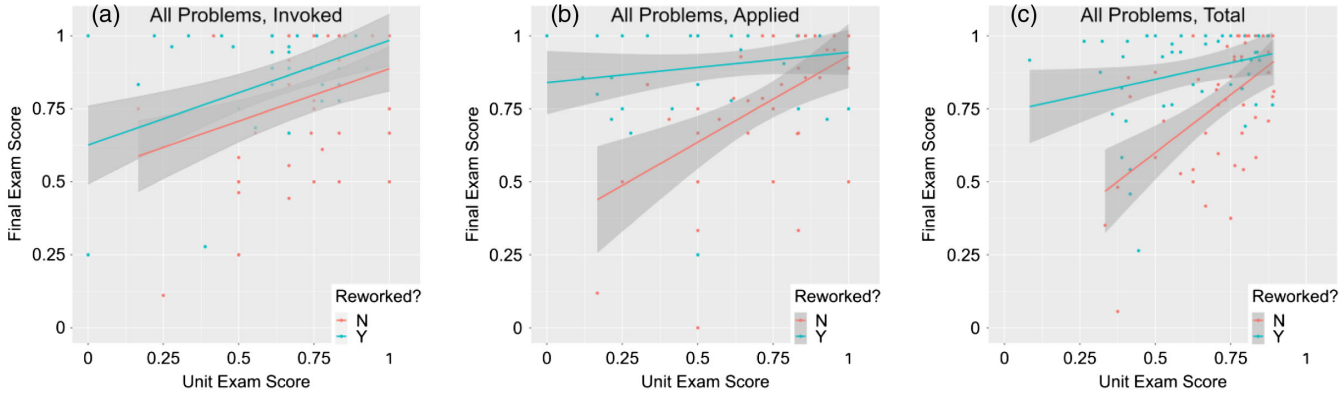


FIG. 3. Summary of results when all problems are aggregated together, for (a) the invoked category, (b) the applied category, and (c) both categories combined into a total score. Discrete data points represent individual student attempts, with blue being those in the did rework population for a given problem and red being those in the did not rework population. For (a) the offset model is depicted, but for (b) and (c) the interaction model provides a significant advantage over the offset model so it is depicted instead.

Instead, the "exam only" serves as a baseline model against which we compare models that account for reworking.

We next consider a model that uses the reworking group as a predictor in addition to unit exam scores. When the regression of final exam scores vs unit exam scores is plotted as in Fig. 3, this model forces the two reworking groups to have the same slope but permits different intercepts [see Fig. 3(a)] We refer to this as the "offset model." Finally, we consider an "interaction model" that allows the two groups to have different relationships between final exam and unit exam scores. When the regression is plotted, this model allows the slope of the regression to vary separately for the two groups [see Figs. 3(b) and 3(c)]. For test A, we assess the overall effect of reworking problems by comparing the "exam only" and offset model. For test B, we compare the offset model to the interaction model to quantify the effects of including an interaction between reworking problems and unit exam scores.

Each test is a likelihood ratio test to compare nested models: the simpler model is a special case of the more complex model. For both tests, the null hypothesis is that the additional term in the model has no effect on the final exam score and the null distribution is a χ^2 distribution with one degree of freedom. For both tests, we report p values derived from the χ^2 distribution as a measure of significance, using the standard definition that p is the probability of obtaining test results at least as extreme as the result actually observed, under the null hypothesis (assumption that a variable has no effect). A very small p value means that an observed outcome is highly unlikely under the null hypothesis that the additional model term has no effect on the final exam score. We use the word "significant" to indicate that we can reject the null hypothesis with a type-I error rate of $\alpha = 0.05$; and "highly significant" to indicate we can reject the null hypothesis with an even more stringent $\alpha = 0.01$. We also report partial η^2 values from ANCOVA to indicate effect size; partial η^2 is the fraction of variance explained by the total variance remaining after accounting for other variables. As is common [35], we use the heuristic that $\eta^2 = 0.01$ indicates a small effect, $\eta^2 = 0.06$ a medium effect, and $\eta^2 = 0.14$ a large effect. For test A, partial η^2 is a direct measure of the effect of reworking the unit exam problem; for test B, it is a measure of the effect of the interaction, i.e., that there is a different dependence between final and unit exam scores for the reworking and nonreworking groups.

F. Interview protocol

To address the third research question, we conducted qualitative analysis in the form of recorded audiovisual interviews with a think-aloud protocol, in order to identify expertlike vs novicelike problem-solving trends among a randomly chosen sample of each of the three sections of EM1. Randomly selected students were recruited to volunteer for interviews after their respective sections were finished, typically within a year's time of completing the course. Table II shows each student, marked by an ID number to protect anonymity, which chose to participate in the interviews. Examples of novicelike [6,14,36,37] and expertlike [3,6,7,32,38–43] thinking, as described in the literature (see Sec. III, part B for details), are marked for each of the interviewed students across each of the three sampled semesters. A pilot interview was conducted with six randomly selected students from the F19 section, which only included P1 and served to finalize decisions on the protocol for future interviews; hence we omit the F19 section from Table II and focus on the S21 and F21 sections with the finalized protocol. Of the randomly selected students from S21 and F21, two responded from the S21 section and four responded from the F21 section. The interview protocol asked students what their current status was (undergraduate, graduate student, employed in the workforce, etc.) and whether they were using the material

TABLE II. A list of interviewed students from S21 and F21 semesters. For the "Rework?" column, a "Y" indicates yes, a "N" indicates no, and an "I" indicates ineligible for pre-post analysis due to scoring at least 90% on the unit exam attempt. For the "Expert-like vs novice-like thinking" column, a dot bullet indicates an expert-like trend and a dash bullet indicates a novice-like trend.

		Rework?				
Term	ID	P1	P2	P3	Expertlike vs novicelike thinking	Current occupation at the time of interview
S21	2	N	Y	Y	• Checking one's work with multiple paths (P1) • Relative improvement with heuristics (P2) - Psyching out exam (P1) - Plug and chug (P3) - Overreliance on formulas (P1, P3)	Has graduated and is working in computer programming. Is preparing for graduate school and studying for the physics GRE (so EM1 is helping to that end)
	11	I	I	Y	• Simplifying picture (P1) • Transfer of learning (P1) • Checking one's work (P2) - Plug and chug (P3) - Overreliance on formulas (P3)	Still undergraduate; is an RA studying physics and applied math (used some of EM1 for EM2, otherwise not)
F21	3	Y	Y	I	• Analogies for transfer of learning (P1) • Cognitive apprenticeship (P1) • Explicit metacognition and reflection (P1, P2) • Working with peers (P2, P3) • Improved self-efficacy (P2) - Psyching out exam (P3) - Plug and chug (P3)	Still undergraduate; researching acoustics (only used EM1 methods in math classes)
	15	I	Y	Y	• Checking mistakes (P1) • Transfer of learning (P1) • Partial improvement (P3) - Drawing inappropriate situation (P2) - Psyching out final exam (P2) - Memorize worked examples (P3)	Just graduated; working on electronics assembly line, planning to return for graduate school (has not used EM1)
	21	Y	I	Y	• Reflection on mistakes (P1) • Effective use of worked examples (P2) • Reworking independently (P2)	Still undergraduate, lab assistant, CS minor with physics major (has used EM1 with electronic devices)
	26	Y	N	N	• Reworking independently (P2) - Memorize worked examples (P1) - Psyching out exam (P3) - Plug and chug (P3)	Just graduated; plans to attend graduate school in fall, focusing on orbital dynamics for research (used EM1 for EM2, also thinks it will be used in grad school)

from EM1 after having finished the course. Students were then presented with a copy of their ungraded unit exam solutions for P1 and asked to explain their work. Once they finished, students were then presented with a copy of their ungraded final exam solutions for P1 (in most cases, described by the interviewer as "a similar problem" but not as the exact same problem) and asked to explain their work on the final exam. The process was repeated for P2 and P3. As an addendum to the protocol to ensure explicit responses, students from the F21 semester were also asked to comment on (a) whether they had felt prepared for the unit exam version of each problem and (b) whether they had seen fit to study their unit exam problems for the final.

The interview protocol questions for the S21 semester were as follows. First, the interviewee's unit exam attempt for P1 was shown, and the interviewee was given the following directions:

(1) This is a problem that you were given to solve on a midterm, and asked to solve. Please walk through your written down solutions and explain to us what your thought process was as you were solving it.
(2) How prepared did you feel for this problem on your unit exam?
(3) How did seeing this problem on the unit test influence your studying for the final, if any?
(4) Now we are going to go over to a very similar problem that was on your final. [The interviewee's final exam attempt is shown here.] Please guide us through your solution, but focus this time on what you did the same and what you did differently this time.

Once the interviewee finished answering the above four questions for P1, the above four questions were repeated for P2 and P3. The interviewee was invited to give any further

comments as desired before the interview ended. After the interview finished, the researchers examined the audiovisual data of the interview for expert-novice behavior. Mentions of types of thinking that could be classified as expertlike or novicelike from cited literature were coded as such upon review of the audiovisual data, as shown in Table II and discussed in Sec. III B.

For P1, four of the six interviewed students were eligible for the quantitative data due to having scored less than 90% on the unit exam attempt; of those four, three students (3, 21, and 26 from F21) chose to rework the problem for credit, while one student (2 from S21) declined. The other two students (11 from S21 and 15 from F21) had scored at least 90% on the unit exam attempt of P1.

For P2, the primarily conceptual problem, four out of six students scored less than 90% on the unit exam attempt and were eligible for the quantitative data; of these, three students (2 from S21; 3 and 21 from F21) chose to rework the problem for partial credit, while one student (26 from F21) declined. Student 11 from S21 and student 15 from F21 scored 90% or more on the unit exam attempt.

For P3, the primarily algorithmic problem, five out of six students (2 and 11 from S21; 15, 21, and 26 from F21) scored less than 90% and were eligible for the quantitative data; four of these students chose to rework the problem, with student 26 from F21 declining to rework the problem. Student 3 from F21 scored more than 90% on the unit exam attempt.

III. RESULTS

A. Quantitative results

Figure 3 is a graphical representation of some of our results, namely when all problems are aggregated together. The offset model is depicted for the invoked category in Fig. 3(a) and is found to be significant albeit with only a small effect size. A larger effect size would be manifested as a larger separation between the blue and red lines. The offset is always positive, which is to say for all three problems the did rework group always scored higher than the did not rework group. In all cases, there is also a positive correlation between unit exam score and final exam score, manifested as a positive slope in the fit lines.

In some cases, however, there is a significant difference in how that correlation manifests itself. For example, in Fig. 3(b), the applied category, the did not rework population has a much stronger dependence on unit exam score (red line, steeper slope) than the did rework population (blue line, shallower slope). For those cases, we employed the interaction model, which allows slopes to vary independently to account for the interaction with the confounding variable. Qualitatively, the effect size for the interaction model is manifested as a large difference in the slope between blue and red lines in Figs. 3(b) and 3(c). Adding the interaction term was found to not have a statistically significant effect on the data in Fig. 3(a) but was highly significant for the applied and total categories (see Tables III and IV, and respective explanations, below). Hence, we have plotted the offset model in Fig. 3(a) but the more complicated interaction model in Figs. 3(b) and 3(c). Similar figures for the individual problems are found in the Appendix; in each case, we have plotted the offset model unless the interaction is found to be significant, in which case the interaction model is plotted. The complete results of test A, namely allowing the offset model as opposed to not allowing a difference between did rework and did not rework groups, are shown in Table III. The first two columns are for all problems considered together as in Fig. 3; the other columns display results for individual problems separately. Significant p values are indicated with *, highly significant with **; and medium and large effect sizes are bolded. As can be seen in the first two columns, there is a large effect size with a highly significant p value for the applied and total categories, whereas the invoked category displays a significant p value but with only a small effect size [this is what is plotted in Fig. 3(a)]. Thus, we can conclude that reworking problems has a statistically significant effect on final exam performance, with particularly large effects in the applied and total categories.

When the problems are considered independently, P1 displays a statistically significant p value with a moderate effect size in the applied and total categories; P2 is statistically significant for the invoked category and nearly

TABLE III. Results of Test A: ANCOVA for the offset model vs exam-only model, i.e., quantifying the overall effect of reworking problems, while also accounting for the correlation between unit and final exam scores [see, e.g., Fig. 3(a)]. Significant p values ($p < 0.05$) are indicated with *, highly significant ($p < 0.01$) with **; and medium and large effect sizes ($\eta^2 > 0.06$) are bolded.

	All problems		P1		P2		P3	
	p	η^2	p	η^2	p	η^2	p	η^2
Invoked	0.034*	0.054	0.14	0.018	0.029*	**0.12**	0.92	0.001
Applied	0.0002**	**0.17**	0.016*	**0.11**	0.052	**0.14**	0.022*	**0.12**
Total	0.0004**	**0.15**	0.012*	**0.070**	0.040*	**0.13**	0.13	0.048

TABLE IV. Results of Test B: ANCOVA for interaction model vs offset model, i.e., quantifying the way the correlation between unit and final exam scores differs for reworked vs did not rework populations (see e.g., Figs. 3(b) and 3(c)]. Significant p values $p < 0.05$) are indicated with *, highly significant ($p < 0.01$) with **; and medium and large effect sizes ($\eta^2 > 0.06$) are bolded.

	All problems		P1		P2		P3	
	p	η^2	p	η^2	p	η^2	p	η^2
Invoked	0.26	0.015	0.62	0.006	0.094	**0.12**	0.12	**0.10**
Applied	0.007**	**0.094**	0.97	0.000	0.18	**0.081**	0.0009**	**0.39**
Total	0.006**	**0.094**	0.86	0.001	0.15	**0.094**	0.0006**	**0.41**

significant for the applied category and has a moderate-to-large effect size for both categories; and P3 is statistically significant only for the applied category and has a moderate effect size.

The complete results of test B, namely allowing the interaction model as compared to the offset model only, are shown in Table IV. Again, the first two columns are for all problems considered together, whereas the other columns are for individual problems; significant p values ($p < 0.05$) are indicated with *, highly significant ($p < 0.01$) with **; and medium and large effect sizes are bolded. As can be seen in the first two columns for all problems aggregated, there is a highly significant p value for both the applied and otal categories, with a medium effect size; this is what is plotted in Figs. 3(b) and 3(c). On the other hand, the p value for the invoked category is not significant, which is why Fig. 3(a) does not plot the interaction model but rather the offset model. Thus, we can conclude that in some cases, there is a significant difference in the interaction between unit exams and final exams, between the two population groups. We will return to that in the next paragraph. When the problems are considered independently, only P3 displays a statistically significant p value and only for the applied and total categories. It is highly significant, and there is a large effect size.

Of note for the cases where the interaction model was statistically significant, is how much shallower the blue curve is than the red curve, especially in Fig. 3(b) and in Figs. 6(b) and 6(c) (in the Appendix). This means that for the did rework group, the unit exam score mostly stops being a good predictor. Just the fact that a student reworked the problem becomes a much better predictor than how well the student did on the problem on the unit exam. When this happens, we conclude that reworking the problem comes close to bringing all of the students who chose to rework up to the level of the best students who did not choose to rework.

B. Qualitative results

1. P1, the composite problem

Student 3 (F21) expressed multiple recognized traits of expertlike problem-solving approaches: *the use of analogies for transfer of learning* [7] on the unit exam attempt, namely the analogy of elastic potential energy in a stretched spring to electric potential energy between opposite charges; *evidence of fading from the cognitive apprenticeship model* [38] via moving on from the spring analogy to demonstrate more explicit understanding of electric potential energy increase on the final exam attempt; and even *explicit metacognition* [39] during the interview itself, via comparing both problem solution attempts to detect an error on the unit exam attempt.

Student 21 (F21) appeared to have a good conceptual understanding of both parts of P1 on the unit exam, such that reworking the problem helped prepare for the final exam mainly in terms of details. However, student 21 did attempt to correct an error on his final exam attempt, which showed an ability to *reflect on his mistakes* [3] even after the course was over.

Student 26 (F21) appeared to benefit from reworking the problem after initially doing poorly on the unit exam attempt of P1, in which the student admitted he had panicked on the grounds of not immediately knowing already how to do the problem completely—suggesting a novicelike trait of attempting to *memorize worked examples* [6] and repeat them verbatim on the final, as opposed to gaining a deeper understanding for problems with similar concepts [36]. However, student 26 was able to explain what the corrections for the unit exam should be. An improved approach for P1's part a on the final exam was begun, but not finished; the student said he put the unfinished part a aside to possibly get back to later, as he was not sure how to finish, but never did return to finish the problem before turning in the exam. Part b on the final exam attempt was correct for student 26.

Student 2 (S21), who chose not to rework P1, said he was unprepared for the unit exam version of P1's part a, by way of not having needed formulas on the formula sheet to do it and having to improvise an incorrect approach. This suggests *psyching out the exam*, by predicting what topics will and will not be featured on that exam (something that also occurred with quantum mechanics students [14]), as well as *reliance on a formula sheet* [37], to the point of not studying the formulas in advance in order to have them

ready. While the student arrived at the correct answer for P1's part b on the unit exam, it was done by explicit calculations of ΔU—this was an allowed solution path for full points but did not necessarily rely upon conceptual understanding. However, on the final exam, student 2 did demonstrate the *use of multiple paths to check his work* on part a, which is expertlike [40].

The two students who scored at least 90% on P1 on the unit exam attempt both demonstrated mastery of the problem during the interviews. For the unit exam, student 11 (S21) was able to *simplify the picture* by refraining from the change in separation of charges to an equivalent change (namely, moving one charge so that it was a distance of 2d away from the other charge as opposed to d) and thus get the problem correct. Part b's explanation drew from a *transfer of learning* about change in mechanical energy (namely, considering the change in kinetic energy the charges would have just before colliding, once they were allowed to accelerate toward each other). Student 11's final exam attempt showcased a different plan than the unit exam attempt, and both paths led to the correct answer for part a.

Student 15 (F21) discussed what few mistakes she made on the unit exam, namely realizing during the unit exam that her answer for part a did not make sense (which still shows some metacognition via *checking her work to see if it makes sense* [40]) and leaning on the explicit calculation to answer part b. Student 15 seems to have studied the problem for the final exam, though she did not say so explicitly, as her answers for the final exam demonstrated a more explicit conceptual explanation for part b that drew from a *transfer of learning* from impulse-momentum concepts from mechanics (namely that the collision between the two charges would deliver a harder impact, or a greater impulse, on each other).

The student interviews for P1 mostly demonstrated either a strong understanding of how to solve the problem on the unit exam or the ability to improve understanding of how to solve it by the final exam. Since almost all students either chose to rework the problem or demonstrated mastery on the unit exam attempt, it is more useful to examine the qualitative examples of how students turned to expertlike vs novicelike approaches to P1 on either attempt. Student 3, who reworked the problem, demonstrated several expertlike traits in discussing both attempts of P1; other students discussed novicelike approaches in studying for the unit exam more so than in attempting the problem itself but appeared to improve with a chance to review the problem for the final exam, whether reworking the problem for partial credit or not.

2. P2, the conceptual problem

Student 2's (S21) unit exam attempt for P2 demonstrated the use of heuristics [41] about how to appropriately draw electric field lines between the two charges; however, student 2 used relatively few heuristics on the unit exam attempt, and while he remembered more heuristics for the final exam attempt, there were still deficiencies in his solution for the latter. Student 2 did notice a sign error in the unit exam attempt (learning from mistakes) during the interview itself but did not remark on other errors.

Student 3 demonstrated several expertlike traits when discussing both his attempts for P2. He *explicitly redid the problem during the interview* in order to correct an error, showcasing metacognitive skills within the interview itself; while reworking the problem, he discussed the benefit of *working with peers* [42]; and noted having *improved heuristics* on the final exam attempt relative to the unit exam attempt. Thus, reworking the unit exam problem seemed to explicitly help student 3 apply expertlike problem-solving approaches. As an aside, student 3 also addressed a *change in self-efficacy* [43] during the unit exam, stating that his confidence about attempting P2 decreased as he worked on it.

Despite having chosen to rework P2 for partial credit, student 15 discussed novicelike behaviors on both attempts of the problem. The student felt underprepared for the unit exam version of the problem, and while some proper heuristics were demonstrated, they were undone because the student *drew a situation that was inappropriate to the problem* (namely substituting a dipole of equal and opposite charges in for the unequal set of charges for a dipole) in order to treat something more familiar. The student also mentioned that she did not explicitly prepare for this problem on the final exam, hence she struggled on the final exam attempt for P2 as well. This implies that student 15 was still attempting to *psych out the final exam*.

In contrast, while student 26 did not rework P2 for partial credit, he did say he was *prompted by his struggles on the unit exam attempt to study it* for the final exam [3], which shows some evidence that at least some students who did not rework a particular problem for credit would still study that problem for the final anyway. The student's attempt at part b improved from the unit exam to the final exam, but the student took some time to realize this from his own work during the interview itself. Of note is the student's discussing a "lack of artistry" in his responses; while the importance of drawing a diagram is typically regarded in terms of describing the situation accurately rather than its aesthetics, the question is raised whether a perceived lack of aesthetics may mask a legitimate misunderstanding of the situation, whether by a student or by a grader.

Student 11, who already scored over 90% on the unit exam attempt of P2, mentioned being prepared for the final exam attempt despite not having explicitly studied it. This suggests that the student might have psyched out the test, or alternately (perhaps simultaneously) already understood the material from the unit exam to his

satisfaction and felt no need to continue studying it. This student explicitly redid the problem during the interview itself, demonstrating a strong ability to *check his work by redoing the problem*.

Student 21 had *studied worked examples from class* [32] in order to prepare for the unit exam, thus scoring over 90% on P2 on the unit exam attempt. Even so, student 21 also said that he recognized a few minor conceptual misunderstandings from the unit exam attempt, and *studied the problem independently, along with all the problems on all the unit exams,* regardless of the incentive for reworking. This suggests that students 11 and 21 performed well on the unit exam version of P2 due to already having study habits that enabled expertlike approaches for this problem.

3. P3, the algorithmic problem

A persistent novicelike tendency toward "plug-and-chug" *approaches* (specifically trying to find the right formula from the permitted sheet of notes) occurred on unit exam attempts for P3, with students 2 and 11 from S21 and students 3 and 26 from F21 all explicitly naming this approach. Student 2 from S21 stated an explicit reliance upon formulas from his sheet of notes and a poor score on part b regarding the dipole moment was attributed to failure to find the right formula accordingly. After reworking P3, student 2 did improve somewhat conceptually on the final exam attempt, but still struggled with the solution.

While student 11 from S21 did attempt a more formal setup of the proper integral for the monopole on the unit exam, *reliance on provided formulas* was still needed for the dipole and quadrupole portions of P3. Upon reworking the problem, student 11 demonstrated a better command of the formal integration methods used for the monopole and dipole moments. However, a plug-and-chug approach persisted for the quadrupole moment on the final exam approach, with student 11 explaining that the formal integration methods gave the "same thing" as formulas provided from a formula sheet.

Despite scoring 90% on the unit exam attempt, student 3 from F21 admitted not expecting to have to study Legendre polynomials that lay the conceptual groundwork for P3's featured potential moments and relied on a formula sheet and the plug-and-chug approach. This suggests another novicelike approach, namely *trying to "psych out" a unit exam.* That being said, student 3 discussed an expertlike response to P3, despite getting close to 100%, by *working with a peer* [42] to more properly understand Legendre polynomials within P3 after the exam was handed back.

Student 15 from F21 did attempt to *draw from worked examples in class* [32] to study the material for P3 but demonstrated a novicelike pitfall to studying worked examples, namely trying to *memorize entire solutions and repeat them verbatim* on the unit exam. Upon failing to recall the quadrupole moment example, student 15 "made stuff up" in an attempt to get partial credit. However, having chosen to rework this problem, student 15 was better conceptually prepared for the final exam and tried to properly derive expressions from Legendre polynomials, although student 15 only got part A correct in this manner.

The last student who reworked the problem, student 21 from F21, did not directly address his attempt to rework the problem during the interview. Student 21's reworking of the problem did not appear to show any benefit, as the respective unit exam and final exam approaches to P3 were identical.

Student 26 from F21, who declined to rework, also attempted to plug and chug through P3 on the unit exam but showed evidence of *psyching out the test* by stating that he did not expect the quadrupole moment to show up at all on the unit exam. This student made a note that the material was important for the final exam; however, this note was insufficient, as while student 26 demonstrated a better conceptual grasp of P3 on the final exam, there were still deficiencies in the solution attempt.

Overall, there was a suggestion that reworking P3 with expertlike approaches in mind did lead to an improved command of the problem solution attempt, via the statements from students 3 (F21) and 11 (S21), as opposed to student 26 (F21) declining to rework and struggling. That being said, other students who reworked the problem, students 2 (S21) and 15 (F21) still showed signs of struggle on the final exam attempt and did not describe expertlike problem-solving approaches as students 3 and 11 did. Some novicelike tendencies persisted from unit exam attempt to final exam attempt regardless of choice to rework, in particular, the plug-and-chug approach. This appears to suggest that the overly algorithmic nature of P3 appears to convince students to rely on novicelike approaches at face value, to the point where even students who choose to rework the problem for explicit incentive may retain novicelike tendencies.

IV. DISCUSSION

A. Quantitative results

1. Overall effect of reworking on final exam performance

Per Figs. 3(a)–3(c), the data distributions within Figs. 3(b) and 3(c) match the trends suggested by Fig. 2 of Brown *et al.* [16], in which QM students who chose to rework the problems for explicit incentive performed relatively well on the final exam almost universally regardless of unit exam score, while students who declined to rework the problems for partial credit still exhibit a visible strong dependence of final exam score vs unit exam score.

To answer the first research question, there is quantitative evidence that explicit incentive to rework unit exam problems results in overall improvement, independent of

unit exam score for first-semester EM students as well as for first-semester QM students in Brown *et al.* For all problems over the entirety of each EM problem attempt [Fig. 3(c)], as well as for the specific applied portions of the problem attempts [Fig. 3(b)], the benefit of reworking unit exam problems holds robustly. For the specific invoked portions of the problem attempts [Fig. 3(a)], the benefit of reworking problems is also suggested, but is somewhat less statistically robust, with only a weak-to-moderate effect size between groups. This suggests that the help received from reworking the problems for the EM student sample is specifically targeted at applying principles correctly, as opposed to invoking correct concepts in the first place, with a strong enough effect to influence overall final exam performance in a similar trend.

Per ANCOVA regression analysis in Table II, test A compared the offset model (namely, including "reworked" vs "did not rework" as a variable) to the exam model (which does not include reworking as a variable) and confirmed that choice of reworking does have a large and highly significant effect on final exam score, both overall and for applied-only scores. There was also a significant effect for invoked-only scores, but only a weak-to-moderate effect size. Test B compared the interaction model (which allowed for interaction between unit exam score and choice of rework) to the offset model (which ignored this interaction), in order to quantify the way the correlation between the unit and final exam scores differs for reworked vs did not rework populations. For test B, there was a moderate effect size for overall and applied-only scores across all three problems, confirming that unit exam score is much less of a predictor of final exam score for students who chose to rework than students who did not. However, this was not the case for invoked-only scores, with only a small effect size and no statistical significance in test B, suggesting that unit exam score may still be conflated with the benefit of the choice of reworking within the applied portion of the problems.

Students appear to receive relatively more benefit on the applied portions of final exam problem attempts than on the invoked portions of those attempts. This suggests that there is relatively modest help in terms of understanding which principles apply to the three featured problems, but more clear help in terms of applying those principles correctly, due to reworking the unit exam problems. One potential reason may be that, in preparing for exams, students may have a relatively easier time understanding which principles will feature on a given exam than understanding specific applications of those principles in the form of carrying out a planned solution. In that case, the benefit of reworking unit exam problems would have a more dramatic effect on applying principles correctly than on invoking correct principles.

2. Effect on conceptual vs algorithmic problems

Turning attention to the second research question, Tables II and III also indicate the differences between the three individual problems: P1, a partly conceptual and partly algorithmic problem; P2, a primarily conceptual problem; and P3, a primarily algorithmic problem. The objective here was to test the hypothesis that the effect of reworking might depend on the type of problem. This could be, for example, because upper-division EM often involves similar concepts to lower-division EM but with new and more rigorous mathematical techniques.

The results in Table II show that the conceptual problem P2 is the only problem to display statistical significance in the invoked category, likely due to a connection between invoking the correct concepts and the primarily conceptual construction of the problem. By contrast, in the primarily algorithmic problem P3, the effect of reworking appears limited to the applied category. Allowing for the additional interaction model as shown in Table III, P3 has an even higher statistical significance in the applied category, along with a very large effect size; and as can be seen in Fig. 6(b) in the Appendix, for this problem in this category, the lower scoring students from the did rework group did essentially as well as the higher scoring students in either group. Finally, in looking at P1, which is a mixture of conceptual and algorithmic, the applied area has the largest statistical significance and effect size when judging the overall effect of reworking (Table II), and the interaction model provided no additional statistically significant benefits (Table III).

The above results appear to consistently uphold the notion that the Invoked portions of EM problems are connected to conceptual performance and applied portions are connected to algorithmic performance. If this is true, then it provides evidence to the assumption that reworking unit exam problems may help upper-division EM students relatively more on algorithmic applications than on invoking concepts correctly.

B. Qualitative results: Expertlike vs novicelike patterns of responses

The third research question regarded whether patterns of responses from a subsample of students using a think-aloud protocol might shed further light on quantitative results. Qualitative trends of expertlike and novicelike problem-solving study habits were identified that potentially help explain quantitative trends. Some students even demonstrated expertlike habits during the interviews themselves, reworking the problem on the spot in order to check their past results.

P1 was drawn from material that occurred relatively early in the semester, therefore students appeared to have a relatively strong command of the problem during interviews whether they reworked the problem or not. That being said, the experience of student 26 (F21) on both attempts of P1 appeared to indicate a more novicelike approach to studying for that particular unit exam, followed by a more expertlike approach to studying upon reworking the problem. This individual result for student 26 is consistent with the concept of scaffolding being of

assistance to the coaching aspect of the cognitive apprenticeship model. Other students who reworked the problem appeared to already have a strong grasp of P1 while discussing the unit exam attempt, already exhibiting expertlike learning strategies, such that reworking the problem might have induced expertlike strategies, or alternately might have reinforced expertlike study habits that these students already had for the unit exam attempt of P1.

There appeared to be less of an enforced trend for P2, the conceptual problem. While expertlike and novicelike patterns emerged similarly to those of P1, they did not appear to correspond to reworking vs declining to rework. A student who reworked P2 struggled on both the unit exam and final exam attempts with novicelike approaches, for example; another student who declined to rework P2, on the other hand, decided to study the P2 unit exam attempt independently anyway and showed improvement from unit exam to final exam. Trends from the specific sampled students, in terms of individual success on the problem, did not necessarily translate to the general positive effect of reworking for P2 on both the invoked and applied portions of the problem during the final exam.

For the algorithmic problem P3, there was more clear evidence that choosing to rework the unit exam attempt did not necessarily reflect expertlike problem-solving approaches as opposed to novicelike approaches, with two students showing expertlike tendencies and two students still showing novicelike tendencies. There was also a persistence of the novicelike plug-and-chug approach to problem-solving on the unit exam attempt, regardless of unit exam performance, as influenced by the use of a formula sheet that students could prepare in advance and bring to the exam, such that students tended to look for the right formula on the sheet to use with the plug-and-chug approach. The students who reworked the problem tended to mention a revisit of the basis for their formulas, namely Legendre polynomials needed to express electric potential moments in integral form prior to explicit calculation. It seems clear that a heavily-algorithmic problem in upper-division EM such as P3 can still induce novicelike problem-solving strategies, even with an opportunity to learn more expertlike strategies by reworking the problems.

V. CONCLUSION

Similar to previous findings with first-semester quantum mechanics [16], an explicit incentive to rework unit exam problems appears to be an effective use of scaffolding within the cognitive apprenticeship model for upper-division electricity and magnetism. The regression model demonstrates that students who choose to rework mistakes on unit exam problems tend to have a better overall experience on the final exam attempt of those problems, regardless of how well or poorly they did on the unit exam attempt. An ANCOVA treatment, using unit exam as the covariate, also demonstrates that this benefit persists for students across all ranges of unit exam score. In contrast, final exam scores for students who decline to rework their mistakes appear to be more dependent upon unit exam scores, indicating that students who struggle with a unit exam problem and do not pursue explicit help will continue to struggle on the final exam version of that problem.

Qualitative interview analysis demonstrates some evidence that reworking the problems induces more expertlike problem-solving approaches. However, there is still evidence that certain novicelike trends persist, particularly for the algorithm-focused P3, regardless of choice to rework; there is also some evidence that students who decline to rework the problems may yet successfully study the problem on their own. The specific task of reworking exam problems appears to be somewhat more beneficial for applying principles correctly than for invoking the correct principles, though the overall benefit exists for both tasks.

The authors also note the possibility that the benefit might not exist just from the reworking of the problems as has been assumed, but perhaps through some additional confounding variable that is intertwined. Two potential examples are as follows: First, the student population who chose to rework the problems might be naturally more effective at studying for the final exam in other ways. For future work, perhaps additional interviews could help separate those two variables. Second, the students' homemade formula sheets for the final exam may have included fully worked individual unit exam problem solutions. If this correlated with reworking, then potentially students who reworked could have had their final exam scores artificially boosted. While this possibility is not accounted for in our statistical analysis, we conclude this was likely not a common practice since only a single student who was eligible to rework more than one problem got 100% on all reworked problems. Nevertheless, future work to fully eliminate this potential confounding variable could entail providing a uniform formula sheet without any worked solutions to all students for the final exam instead of allowing homemade formula sheets.

Other limitations of this study are evident through the use of exam problems to examine student learning of problem-solving in upper-division electricity and magnetism, as exam problems are just one of several traditional measurements of course performance. The aspect of "algorithmic" vs "conceptual" framing of problem questions may be more thoroughly explored in future research with exam problems that each have a conceptual portion and an algorithmic portion; this option was not used for this particular study since the potential differences between algorithmic and conceptual problems were not recognized until after the study was complete. Furthermore, it is necessary to develop and test more explicit scaffolding materials, e.g., tutorials, that are more explicitly based upon

a theoretical framework, such as a prescribed problem-solving framework as coached within the cognitive apprenticeship model. Current literature features tutorials for upper-division courses focused on conceptual understanding, which is one of several topics that must be considered for overall problem-solving effectiveness in designing future instruments. A more complete treatment of a problem-solving framework, while arguably more complex, may provide further benefit to improving student problem-solving skills at the upper-division level.

ACKNOWLEDGMENTS

The authors would like to thank Chandralekha Singh for her advice and assistance in beginning this collaboration, as well as her contributions to validating the rubrics; Robert P. Devaty for his contributions to validating the rubrics; and Daniel Jones, Isa Kohls, Ethan Edwards, and Logan Page for their roles in grading and interrater reliability.

APPENDIX: QUANTITATIVE RESULTS PLOTTED FOR EACH PROBLEM SEPARATELY

This appendix includes further graphs of linear regression models similar to those of Figs. 3(a)–3(c), specifically looking at each of the three problems separately: Fig. 4 features P1, Fig. 5 features P2, and Fig. 6 features P3, in the same manner as Fig. 3 represents data for all three problems together. Offset models require "did rework" and "did not rework" groups to have the same slope in order to illustrate the differences due to reworking via a shift in y-intercept. These models are used in the plots, except when Interaction models, in which treatment and comparison groups are allowed different slopes, indicate statistical significance.

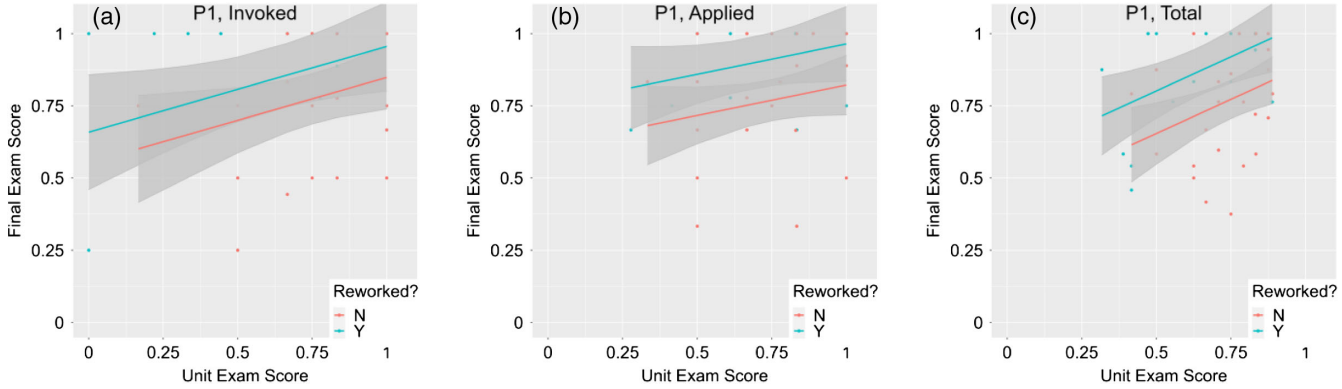


FIG. 4. Summary of results for P1 alone, for (a) the invoked category, (b) the applied category, and (c) both categories combined. The offset model was used for all three.

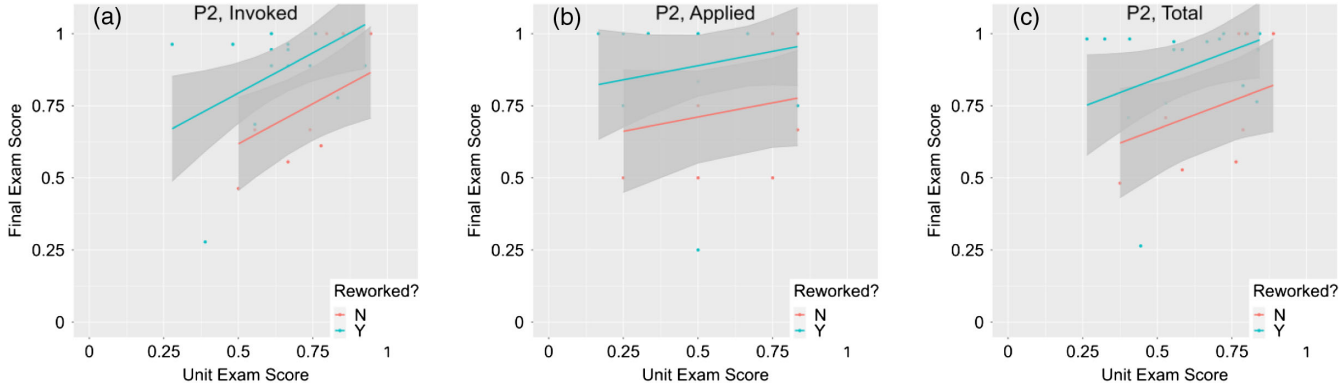


FIG. 5. Summary of results for P2 alone, for (a) the invoked category, (b) the applied category, and (c) both categories combined. The offset model was used for all three.

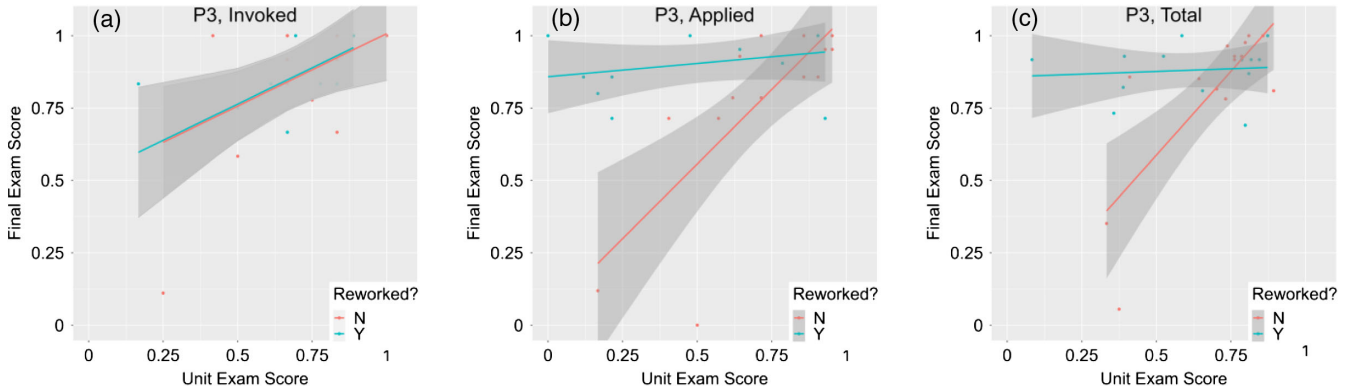


FIG. 6. Summary of results for P3 alone, for (a) the invoked category, (b) the applied category, and (c) both categories combined. In this case, the interaction model (different slopes for the two populations) provided a significant advantage over the offset model in the applied and total categories and is used in (b) and (c), but for invoked in (a) the offset model was used.

[1] L. Hsu, E. Brewe, T. M. Foster, and K. A. Harper, Resource letter RPS-1: Research in problem solving, Am. J. Phys. **72**, 1147 (2004).

[2] L. Bao and K. Koenig, Physics education research for 21st century learning, Discip. Interdiscip. Sci. Educ. Res. **1**, 2 (2019).

[3] E. Yerushalmi, E. Cohen, A. Mason, and C. Singh, What do students do when asked to diagnose their mistakes? Does it help them? I. An atypical quiz context, Phys. Rev. ST Phys. Educ. Res. **8**, 020109 (2012); E. Yerushalmi, E. Cohen, A. Mason, and C. Singh, What do students do when asked to diagnose their mistakes? Does it help them? II. A more typical quiz context, Phys. Rev. ST Phys. Educ. Res. **8**, 020110 (2012).

[4] e.g., P. Klaczynski, Metacognition, and cognitive variability: A dual-process model of decision making, and its development, in *The Development of Judgment, and Decision Making in Children, and Adolescents*, edited by J. Jacobs and P. Klaczynski (L. Erlbaum Associates, Mahwah, NJ, 2005), pp. 39–76; J. Speirs, M. Stetzer, B. Lindsey, and M. Kryjevskaia, Exploring and supporting student reasoning in physics by leveraging dual-process theories of reasoning and decision making, Phys. Rev. Phys. Educ. Res. **17**, 020137 (2021).

[5] e.g., A. Maries and C. Singh, Do students benefit from drawing productive diagrams themselves while solving introductory physics problems? The case of two electrostatics problems, Eur. J. Phys. **39**, 015703 (2017); A. Maries and C. Singh, Case of two electrostatics problems: Can providing a diagram adversely impact introductory physics students' problem solving performance?, Phys. Rev. Phys. Educ. Res. **14**, 010114 (2018).

[6] e.g., E. Redish, J. Saul, and R. Steinberg, Student expectations in introductory physics, Am. J. Phys. **66**, 212 (1998); W. Adams, K. Perkins, N. Podolefsky, M. Dubson, N. Finkelstein, and C. Wieman, New instrument for measuring student beliefs about physics and learning physics: The Colorado Learning Attitudes about Science Survey, Phys. Rev. ST Phys. Educ. Res. **2**, 010101 (2006); A. Mason and C. Singh, Surveying college introductory physics students' attitudes and approaches to problem solving, Eur. J. Phys. **37**, 055704 (2016).

[7] e.g., C. Singh, Transfer of learning in quantum mechanics, AIP Conf. Proc. **790**, 23 (2005); *Transfer of Learning from a Modern Multidisciplinary Perspective*, edited by J. Mestre (Information Age Publishing, Greenwich, CT, 2006); "Transfer of learning in problem solving in the context of mathematics and physics." In *Learning to Solve Complex Scientific Problems*, edited by D. H. Jonassen (Routledge, New York, 2007), pp. 223–246.

[8] e.g., J. Barajas, Problem solving, and writing I: The point of view of physics, Lat. Am. J. Phys. Educ. **1**, 4 (2007), http://lajpe.org/sep07/BAROJAS_Final.pdf; W. B. Lane, Letters home as an alternative to lab reports, Phys. Teach. **52**, 397 (2014); J. Hoehn and H. Lewandowski, Incorporating writing in advanced lab projects: A multiple case-study analysis, Phys. Rev. Phys. Educ. Res. **16**, 020161 (2020).

[9] e.g., P. Heller, R. Keith, and S. Anderson. Teaching problem solving through cooperative grouping. Part 1: Group versus individual problem solving, Am. J. Phys. **60**, 627 (1992).

[10] P. Heller and M. Hollabaugh, Teaching problem solving through cooperative grouping. Part 2: Designing problems and structuring groups, Am. J. Phys. **60**, 637 (1992).

[11] P. Antonenko, C. Ogilvie, D. Niederhauser, J. Jackman, P. Kumsaikaew, R. Marathe, and S. Ryan, Understanding student pathways in context-rich problems, Educ. Inf. Technol. **16**, 323 (2011).

[12] e.g., R. Sayer, A. Maries, and C. Singh, Quantum interactive learning tutorial on the double-slit experiment to improve student understanding of quantum mechanics, Phys. Rev.

Phys. Educ. Res. **13**, 010123 (2017); E. Marshman and C. Singh, Investigating and improving student understanding of the expectation values of observables in quantum mechanics, Eur. J. Phys. **38**, 045701 (2017).

[13] e.g., W. Christensen, D. Meltzer, and C. Ogilvie, Student ideas regarding entropy and the second law of thermodynamics in an introductory physics course, Am. J. Phys. **77**, 907 (2009); T. Smith, J. Thompson, and D. Mountcastle, Student understanding of Taylor series expansions in statistical mechanics, Phys. Rev. ST Phys. Educ. Res. **9**, 020110 (2013).

[14] A. Mason and C. Singh, Do advanced physics students learn from their mistakes without explicit intervention?, Am. J. Phys. **78**, 760 (2010).

[15] A. Mason and C. Singh, Surveying graduate students' attitudes and approaches to problem solving, Phys. Rev. ST Phys. Educ. Res. **6**, 020124 (2010).

[16] B. Brown, A. Mason, and C. Singh, Improving performance in quantum mechanics with explicit incentives to correct mistakes, Phys. Rev. Phys. Educ. Res. **12**, 010121 (2016).

[17] e.g., J. P. Zwolak and C. A. Manogue, Assessing student reasoning in upper-division electricity and magnetism at Oregon State University, Phys. Rev. ST Phys. Educ. Res. **11**, 020125 (2015); B. Wilcox and S. J. Pollock, Validation and analysis of the coupled multiple response Colorado upper-division electrostatics diagnostic, Phys. Rev. ST Phys. Educ. Res. **11**, 020130 (2015); C. Baily, Q. X. Ryan, C. Astofoli, and S. J. Pollock, Conceptual assessment tool for advanced undergraduate electrodynamics, Phys. Rev. Phys. Educ. Res. **13**, 020113 (2017); B. Xue, Identifying and addressing student difficulties with the source-field relationship using tutorials in upper-division electromagnetism, Ph.D. dissertation, University of Washington, 2021.

[18] E. Marshman, A. Maries, R. Sayer, C. Henderson, E. Yerushalmi, and C. Singh, Physics postgraduate teaching assistants' grading approaches: Conflicting goals and practices, Eur. J. Phys. **41**, 055701 (2020).

[19] A. Mason and J. Colton, Reworking exam problems to incentivize improved performance in upper-division electrodynamics, presented at PER Conf. 2022, Grand Rapids, MI, 10.1119/perc.2022.pr.Mason.

[20] e.g., C. Henderson and K. Harper, Quiz corrections: Improving learning by encouraging students to reflect on their mistakes, Phys. Teach. **47**, 581 (2009).

[21] e.g., L. Deslauriers, E. Schelew, and C. Wieman, Improved learning in a large-enrollment physics class, Science **332**, 862 (2011).

[22] M. C. Wittmann, R. N. Steinberg, and E. F. Redish, Investigating student understanding of quantum physics: Spontaneous models of conductivity, Am. J. Phys. **70**, 218 (2002).

[23] L. Bao and E. F. Redish, Understanding probabilistic interpretations of physical systems: A prerequisite to learning quantum mechanics, Am. J. Phys. **70**, 210 (2002).

[24] B. Modir, J. Thompson, and E. Sayre, Students' epistemological framing in quantum mechanics problem solving, Phys. Rev. Phys. Educ. Res. **13**, 020108 (2017).

[25] B. Wilcox, M. D. Caballero, D. Rehn, and S. Pollock, Analytic framework for students' use of mathematics in upper-division physics, Phys. Rev. ST Phys. Educ. Res. **9**, 020119 (2013).

[26] M. D. Caballero and S. Pollock, A model for integrating computation without changing the course: An example from middle-division classical mechanics, Am. J. Phys. **82**, 231 (2014).

[27] e.g., G. Polya and J. Conway, *How to Solve It: A New Aspect of Mathematical Method (Vol. 2)* (Princeton University Press, Princeton, NJ, 1957); A. H. Schoenfeld, Pólya, problem solving, and education, Math. Mag. **60**, 283 (1987); J. Docktor, J. Dornfeld, E. Frodermann, K. Heller, L. Hsu, K. A. Jackson, A. Mason, Q. Ryan, and J. Yang, Assessing student written problem solutions: A problem-solving rubric with application to introductory physics, Phys. Rev. Phys. Educ. Res. **12**, 010130 (2016).

[28] e.g., P. Emigh and C. Manogue, Finding derivatives from an equipotential graph, presented at PER Conf. 2022, Grand Rapids, MI, 10.1119/perc.2022.pr.Emigh.

[29] E. Marshman and C. Singh, Investigating and improving student understanding of quantum mechanical observables and their corresponding operators in Dirac notation, Eur. J. Phys. **39**, 015707 (2017); E. Marshman and C. Singh, Validation and administration of a conceptual survey on the formalism and postulates of quantum mechanics, Phys. Rev. Phys. Educ. Res. **15**, 020128 (2019).

[30] J. Sweller, Cognitive load during problem solving: Effects on learning, Cogn. Sci. **12**, 257 (1988); J. Sweller, Cognitive load theory: Recent theoretical advances, in *Cognitive Load Theory*, edited by J. L. Plass, R. Moreno, and R. Brünken (Cambridge University Press, New York, 2010), pp. 29–47.

[31] e.g., A. Maries, S.-Y. Lin, and C. Singh, Challenges in designing appropriate scaffolding to improve students' representational consistency: The case of a Gauss's law problem, Phys. Rev. Phys. Educ. Res. **13**, 020103 (2017).

[32] A. H. Schoenfeld, Problem solving in context(s), in *The Teaching, and Assessing of Mathematical Problem Solving*, edited by R. I. Charles and E. A. Sliver (Lawrence Erlbaum Associates, The National Council of Teachers of Mathematics, Inc., Reston, VA, 1988), Vol. 3, pp. 82–92; e.g., E. Harskamp and C. Suhre, Schoenfeld's problem solving theory in a student controlled learning environment, Comput. Educ. **49**, 822 (2007).

[33] P. Black and D. Wiliam, Developing the theory of formative assessment, Educ. Asse. Eval. Acc. **21**, 5 (2009); R. Dufresne and W. Gerace, Assessing-to-learn: Formative assessment in physics instruction, Phys. Teach. **42**, 428 (2004).

[34] D. Griffiths, *Introduction to Electrodynamics*, 4th ed. (Cambridge University Press, New York, 2017), 599 pages. The instructor also allowed the used 3rd edition if students preferred.

[35] J. Miles and M. Shevlin, *Applying Regression and Correlation: A Guide for Students and Researchers* (Sage Publications Ltd., London, 2001).

[36] A. Elby, Helping physics students learn how to learn, Am. J. Phys. **69**, S54 (2001).

[37] e.g., J. Mestre, R. Dufresne, W. Gerace, P. Hardiman, and J. Touger, Promoting skilled problem-solving behavior among beginning physics students, J. Res. Sci. Teach.

30, 303 (1993); D. Rehfuss, Formula sheet caveat, Phys. Teach. **41**, 375 (2003).

[38] A. Collins, J. Brown, and S. Newman, Cognitive apprenticeship: Teaching the crafts of reading, writing, and mathematics, in *Knowing, Learning, and Instruction: Essays in Honor of Robert Glaser*, edited by L. Resnick (Routledge, New York, 1989), pp. 453–494.

[39] J. Flavell, Metacognitive and cognitive monitoring: A new area of cognitive developmental inquiry, Am. Psychol. **34**, 906 (1979).

[40] A. Yimer and N. Ellerton, Cognitive and metacognitive aspects of mathematical problem solving: An emerging model, in *Identities, Cultures and Learning Spaces: Proceedings of the 29th Annual Conference of the Mathematics Education Research Group of Australasia, Canberra, ACT, Australia, 2006*, edited by P. Grootenboer, R. Zevenbergen, and M. Chinnappan (MERGA, Sydney, Australia, 2006), pp. 575–582.

[41] F. Reif, J. Larkin, and G. Brackett, Teaching general learning and problem-solving skills, Am. J. Phys. **44**, 212 (1976).

[42] A. Mason and C. Singh, Helping students learn effective problem solving strategies by reflecting with peers, Am. J. Phys. **78**, 748 (2010).

[43] V. Sawtelle, E. Brewe, R. M. Goertzen, and L. Kramer, Identifying events that impact self-efficacy in physics learning, Phys. Rev. ST Phys. Educ. Res. **8**, 020111 (2012).