Impact of teaching students to use evaluation strategies

Aaron R. Warren
Department of Mathematics/Statistics/Physics, Purdue University North Central, 1401 S. US-421, Westville, Indiana 46391, USA
(Received 30 January 2010; published 23 July 2010)

Students often make mistakes in physics courses and are expected to identify, correct, and learn from their mistakes, usually with some assistance from an instructor, textbook, or fellow students. This aid may come in many forms, such as problem solutions that are given to a class, tutoring to an individual student, or a peer discussion among several students. However, in each case a student relies upon an external agent in order to determine whether, and how, her work is mistaken. Consequently, the student's learning process is largely contingent upon the availability and quality of external evaluating agents. One may suspect that if a student developed the ability to evaluate her own work, her dependence on external agents could be moderated and result in an enhancement of her learning. This paper presents the results of a study investigating the impact of novel activities that aim to teach students when, why, and how to use the strategies of unit analysis and special-case analysis. The data indicate that it is possible to help students dramatically improve their understanding of each strategy, and that this has a significant impact on problem-solving performance.

DOI: 10.1103/PhysRevSTPER.6.020103 PACS number(s): 01.40.Fk

I. INTRODUCTION

Many students are almost completely reliant upon external evaluators in order to check and correct their work. Physics courses are generally structured so that students receive feedback from instructors, peers, or intelligent tutoring systems which enable the identification and correction of mistakes in the students' problem solutions. Since no or little explicit attention is paid to helping students develop the means with which to evaluate their own work, it is natural for the students to develop a belief that evaluation by external authorities is the only way to identify and learn from their mistakes. We begin by outlining a pair of strategies students can use to internally evaluate their own work, and which may thereby enable some degree of self-regulated learning.

II. EVALUATION STRATEGIES

The ability to effectively evaluate information has long been recognized as an important cognitive process and educational objective [1,2]. According to Anderson & Krafwohl, "Evaluation is defined as making judgments based on criteria and standards," (p. 83). In general, a given particular (i.e., piece of information) is evaluated by determining whether it satisfies some set of criteria to such a degree as to pass a pre-established standard. In physics, there are several types of criteria and standards, and general guidelines as to how the criteria and standards are to be applied. These associations of criteria, standards, and methods of application, are called *evaluation strategies*.

There are many evaluation strategies for different types of particulars in physics. For example, a proposed problem-solution may be evaluated using the strategy of unit analysis to check whether each equation in the solution is physically sensible. A proposed theoretical model can be evaluated by checking whether it is consistent with other models in certain limiting cases. An experimental result can be evaluated by developing an independent experimental method and checking whether the two experiments give consistent results.

The research reported here is guided by a belief that evaluation strategies play a critical role in developing a coherent, hierarchally organized, robust working knowledge of physics. Before discussing why it may be important for students to learn and use evaluation strategies, two such strategies are outlined below. Each of these strategies could be cast in the form of hypothetico-deductive reasoning [3,4], whereby a hypothesis regarding the given particular is first constructed and then tested. The type of information each strategy is meant to evaluate and the criteria by which the information is judged are summarized below.

A. Unit analysis

Unit analysis is used to evaluate equations to determine whether they are physically sensible, having the same units for each term. If the equation is found to have inconsistent units on some terms, there are three possible reasons: (a) the equation is incorrect; (b) the student incorrectly remembers the units for some quantities; (c) the student made an algebraic mistake in her analysis.

Ideally, if the equation is incorrect the student should determine exactly how the equation fails the unit analysis and to then figure out which quantities and operations need to be added or removed in order to satisfy the unit analysis. In this way the student can, in principle, correct the equation to make it physically coherent.

B. Special-case analysis

Special case analysis is used to evaluate an equation, model, or conceptual claim to determine whether it is consistent with prior knowledge and experience. Any equation, model, or claim is meant to be true for some range of physical situations. This strategy, as defined for the purposes of this study, requires the students to choose some specific situation of which they have prior knowledge, and which also lies within or at the limits of the range of applicability. They

then determine what the equation, model, or claim predicts for this situation, and compare the prediction with our knowledge of what actually happens. Essentially, the strategy is to conduct a thought experiment determining whether the equation, model, or claim makes sense based on prior knowledge.

If the equation, model, or claim is found to be inconsistent with prior knowledge, there are three possible reasons: (a) the equation, model, or claim is incorrect; (b) the student's prior knowledge of the situation is incorrect; (c) the student made a mistake in the execution of the analysis.

If the equation, model, or claim is incorrect, the student should determine exactly how the equation, model, or claim fails the special-case analysis. The goal is to figure out how the equation, model, or claim needs to be modified in order to make it consistent with the prior knowledge. While this strategy is quite general, the actual process is highly dependent on the details of the physical model being employed, and the particular context of the problem.

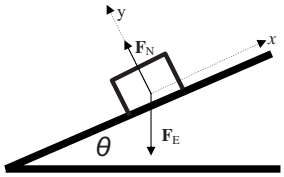
III. REMARKS

It is important to note that each evaluation strategy is subject to errors, as it is wholly possible for the student to make a mistake while using these strategies. Therefore the evaluation strategies listed above provide a useful though imperfect means for judging the validity and soundness of proposed solutions to quantitative questions. Also, there are certainly many other evaluation strategies used in physics, such as error analysis to evaluate an experimental result [5]. Research has found that it is possible and beneficial to help students adopt certain paradigms and strategies for evaluating experimental data [6,7]. One of the major instructional challenges identified by Lippmann was helping students adopt a frame which values the theory of measurement, as many students were largely unaware that such a theory exists and is essential to good science. Based on this, one may fully expect that one of the difficulties in helping students learn the evaluation strategies of unit and special-case analysis will be helping the students to adopt a frame which values these strategies.

IV. EVALUATION AND LEARNING

Evaluation strategies may serve a role in regulating the structure of schemas employed by students in answering questions [8–11]. That is, an evaluation strategy may be linked to an array of schemata responsible for context-specific activities, such as solving inclined-plane problems. For the purposes of this paper, a schema is defined to be an interconnected network of knowledge elements that can be triggered by a variety of inputs, which attempts to organize and assign meaning to elements of the input, and then utilize the input to produce some set of cognitive outputs. When a context-specific schema is activated, an evaluation strategy has some probability of being activated as well. The use of an evaluation strategy can lead the student to recognize that the problem-solving schema is incoherently structured or gives results that are inconsistent with the results of another

Question: What is the normal force exerted on a block of mass M by a surface which is inclined at an angle θ above the horizontal?



Student solution:

X: $-Mg\ cos(\theta) = M\ a_x$ (after substituting $F_E = Mg$)

Y: $F_N - Mg\ sin(\theta) = M\ a_y = 0$ N

Therefore,

$F_N = Mg\ sin(\theta)$

FIG. 1. An illustrative example of a traditional homework problem and a possible student solution.

schema. Upon such recognition, the student may correct her own mistake, consequently restructuring the associated schema. In other words, it is possible that evaluation strategies may be one of the agents responsible for establishing *local* and *global coherence* [12], and also for modifying the conditions under which a particular schema is activated.

As a simple illustrative example, consider a student working on the homework problem shown in Fig. 1. The student's attempted solution is also shown. In this case, the student's work is fine except for the incorrect use of trigonometry when determining the force components. It may be that the student used *cos* and *sin* as she did because she has a strong schematic connection that associates cos with the x axis, and sin with the y axis. By doing a special-case analysis of her solution, say by examining the case when $\theta=0°$, the student may realize that she has made a mistake. If $\theta=0°$, then we expect $F_N=Mg$ since the incline will be level, requiring the normal force to completely balance the gravitational force. However, plugging $\theta=0°$ into the student's solution gives $F_N=0$ N. This analysis therefore indicates that the student made a mistake in the way she dealt with the angle θ in her solution.

If the student tries the simplest alternative by swapping the sin and cos functions in her solution, she will get a new solution that does pass this special-case analysis. This process of evaluation and self-correction could thereby alter the student's schema, as she may in the future be less likely to blindly associate cos with the x axis and sin with the y axis. In particular, a successful special-case analysis would elaborate on why cos should be associated with the y axis, and sin with the x axis in this case. The links between knowledge elements in the student's mind may consequently be revised to account for this surprising result. Although there are many steps where a student may become lost or make a mistake, it is possible that the use of evaluation strategies can serve a regulatory function in the organization of student knowledge.

Without this evaluation strategy, a student's ability to revise and regulate her understanding is much more limited. If the sole means for evaluation and feedback lie outside our students, they become dependant upon external agents in order to engage in learning. This dependence strongly constrains the range of times and places in which a student has the opportunity to learn, as the majority of student learning can only occur with the feedback of an instructor, textbook,

or peer. In many college-level courses, especially large-enrollment courses, each student has only a few hours each week in the presence of an instructor, and only during a fraction of that time does an instructor directly interact with each student. Textbooks fail to provide any sort of dynamic feedback for students, and peer feedback may be limited by a lack of availability (and often may be incorrect or misleading). Consequently, it should come as no surprise that students often learn far too little in physics courses, although they can learn more when courses are structured to facilitate better evaluative feedback to students via interactive-engagement methods [13,14], such as implementations of the University of Washington Tutorials [15] at the University of Colorado [16], Peer Instruction [17,18], and intelligent tutoring systems [19,20]. In each case, a student is given more extensive access to finely-tuned external evaluation, and hence has a greater opportunity to learn.

If students had the ability to evaluate their own work, their learning would no longer be completely contingent upon external evaluators. Instead, students would be able to identify and correct their own mistakes, allowing them to better learn on their own. In fact, Zimmerman and Martinez-Pons identified self-evaluation as a necessary component in their model of self-regulated learning [21], and self-evaluation has been shown to promote self-regulated learning in young students [22]. Likewise, Hammer [23] has identified the importance of student "independence" which typifies the extent to which a student takes responsibility for constructing their own knowledge (instead of simply accepting what is given by authorities without any evaluation).

An important potential benefit of student self-evaluation is that it may help students develop authentic science reasoning abilities, a major goal of science education [24–26]. In particular, the use of evaluation strategies highlights the fact that our judgments of theoretical models can produce false positives and negatives, and this limits our confidence in such judgments. Recognition of the sources of reasoning errors is fundamental to many aspects of critical thinking [27] and reflective judgment [28]. It seems reasonable to believe, then, that these abilities may be enhanced if instructors can help students to value and use evaluation strategies.

V. TEACHING AND ASSESSING EVALUATION

To help students learn how and why to use evaluation strategies, a set of formative activities and rubrics were developed [29,30]. In general, students must understand the goal state they are trying to achieve, their current level of performance, and how to utilize descriptive feedback to improve their performance. A scoring rubric is one tool that can be used to help achieve these conditions. The rubrics break up abilities into finer-grained component abilities, and contain descriptions of different levels of performance, including the target level. A student or a group of students can use the rubrics to self-assess her or their own work, or an instructor can use the rubrics to assess students' responses and provide feedback [31–33]. Additionally, written and verbal comments on student work may be given to personalize the feedback and address particular issues in a student's work.

VI. EVALUATION TASKS AND RUBRICS

During the 2002–2003 and 2003–2004 academic years a library of tasks were designed, tested, and refined to help students learn the evaluation strategies of unit analysis and special-case analysis. Below is a list of the types of tasks featured in this study. An example packet of these and other types of evaluation tasks can be downloaded from http://paer.rutgers.edu/ScientificAbilities/Kits/default.aspx.

External unit analysis—A problem and proposed solution are given, and the student must do a unit analysis to evaluate (and possibly revise) the given solution.

External special-case analysis—A problem and proposed solution are given, and the student must do a special-case analysis to evaluate (and possibly revise) the solution.

Conceptual counterexample—a conceptual claim is made, and the student is asked whether they agree or disagree, and to justify their opinion. In many cases, the most appropriate strategy is to do a special-case analysis, although there is no specific prompting to use any particular strategy.

Integrated tasks—The student must solve a problem, then do unit and/or special-case analysis to evaluate their solution (and possibly revise it).

Critical thinking tasks—The student is given a problem and proposed solution, and asked to give three independent arguments which each analyze whether the given solution is reasonable. There are no explicit prompts to use any particular strategy, so the student must spontaneously recognize that special-case and unit analysis can each be used to generate arguments, and then use these strategies to make valid and sound arguments.

These tasks may be used in recitation or homework assignments. Using the scoring rubrics and giving descriptive comments on student work for evaluation tasks provides formative assessment of the students' work. The scoring rubrics rate student performance on a scale of 0–3, with the following general meanings; 0=no meaningful work done, 1=student attempts but does not understand the general method for completing the task, 2=student understands the general method, but her execution is flawed, 3=student's method and execution of the task are satisfactory. The process by which we developed the rubrics is discussed in Etkina *et al.* [34]

A rubric score of 2 indicates the student tried applying all the steps of the special-case analysis (SCA), but made some superficial mathematical or conceptual error along the way. In other words, the student demonstrates that she understands the context-independent, strategic reasoning process of SCA, but commits some minor mistake in using the context-specific information during the application of the SCA (e.g., forgetting a minus sign, or confusing x and y components of a vector). So the difference between a 2 and a 3 is only indicative of how well the student used the particular context-specific information in the task; to get either score, the student must demonstrate sound use of the general SCA reasoning process. Our final set of evaluation rubrics yielded an average inter-rater agreement level of 96%, ranging between 91% and 99%, and a Cohen Kappa [35] of 0.947 ($p < .001$). The rubrics are available for download at http://paer.rutgers.edu/ScientificAbilities/Rubrics/default.aspx, as Rubric I.

VII. STUDY DESIGN AND RESULTS

Preliminary research conducted during the 2003–2004 academic year demonstrated significant correlations between students' abilities to use evaluative strategies and their problem-solving performance [36]. During the 2004–2005 academic year, a study was conducted to further investigate the effects of using evaluation tasks in an algebra-based physics course. After discussing the design of the study, results are presented addressing three research goals: (1) measure students' abilities to use evaluation strategies; (2) investigate the extent to which students valued and incorporated evaluation strategies into their personal learning behavior; (3) test whether the use of evaluation tasks resulted in significant improvements in student problem-solving performance.

To accomplish these goals, data were collected from two courses at Rutgers University as part of a quasiexperimental control group study. The experiment group consists of the 193–194 course, a year-long introductory algebra-based physics course for life science majors. The gender distribution was 46% male, 54% female. There were 200 students for each term, with 95% of students participating in both the 193 and 194 courses. The comparison group is the year-long 203–204 course, another introductory algebra-based course for life science majors. There were 459 students for the fall term and 418 of those students continued enrollment in the spring term. These two courses were generally run in parallel, although the 203–204 class covers slightly more material and the students in this class typically have stronger math and science backgrounds than in the 193–204 course (it is the more competitive course, held on a different university campus). This likely bias was substantiated by the data and will serve to accentuate the results of our study, as discussed below.

The lectures for 203–204 were all designed and given by Alan Van Heuvelen. The lectures for the 193 course were given by Sahana Murthy (a post-doctoral student working with the Rutgers Physics Astronomy & Education Group, at the time), and the lectures for the 194 course were given by the author. All lectures for 193 and 194 were based on Van Heuvelen's lecture notes. Lectures for all courses involved chalkboard and transparency-based presentations featuring experimental demonstrations and student engagement via peer discussions and student infrared response systems. The chalkboard and transparency-based presentations often began with an experimental demonstration to establish a concrete example of some new class of phenomena to be studied. Questions about the properties of the phenomena would be elicited and/or posed by the lecturer. These questions would motivate the construction of a physical model. Application of the model to solve conceptual and traditional problems would then be illustrated. This entailed making multiple representations of information, assessing consistency between the different representations, and utilizing the representations together with the physical model in order to solve the problem. Students would then work in groups on one or two problems, respond via infrared response systems, and receive some feedback from the lecturer. It should be noted that lectures in the 203–204 courses did include explicit modeling of unit analysis. One premise of the study was the ability to provide very similar lecture environments for the two courses. However, some stylistic and personal differences in lecturing were unavoidable, and the threat to the study's internal validity will be assessed below.

Recitations for both courses involved ~25 students working in groups of 3–5 on a recitation assignment, with a teaching assistant there to provide help as needed. Homework assignments featured problems which were distinct but usually similar to problems from the recitation assignments. Solutions for recitation and homework assignments were posted online for each course. The recitation and homework assignments given by Van Heuvelen for 203–204 were generally identical to those given in 193–194 except that some problems from the 203–204 assignments were replaced by evaluation activities. In general, there were between one and two such replacements each week, with one replacement in the homework assignment and usually one replacement in the recitation assignment.

This replacement served as our experimental factor, being the only designed difference between the two courses. We attempted to minimize any potential bias due to time-on-topic differences by only replacing problems which covered the same material, and which were thought to take roughly the same amount of time to complete, as the evaluation tasks which were used in their place. Some differences in the recitation and homework assignments were inevitable as the 203–204 course is required to cover more material, and had 55-min recitations while the 193–194 course had 80-min recitations. Again, we will address the threats posed by such factors later.

The evaluation tasks used in recitations and homework for 193–194 covered only half of the topics in the course. For example, we did not include any evaluation tasks relating to momentum, fluid mechanics, or wave optics, although tasks were included on work-energy, the first law of thermodynamics, and dc circuits (see Warren [37] for a list of specific evaluation tasks used). This is an important feature of the study since it provides a baseline response, which will be useful when testing whether the use of evaluation tasks affects problem-solving performance.

Laboratories for both courses were practically identical, with ~25 students per section working in groups of 3–5 with the aid of a teaching assistant. The laboratories featured three types of experiments; investigative, testing, and application. Each laboratory usually included two experiments, most of which were either testing or application experiments. It should be noted that although the 193–194 students were required to take the laboratory, the 203–204 students were not, as the laboratories constitute a distinct course (labeled 205–206). However, nearly all (~98%) of the 203–204 students were enrolled in 205–206 concurrently.

Another important set of factors in the study are the teachers themselves. The teaching assistants in 193–194 were atypical, as 4 of the 8 assistants for the course were enrolled at the Graduate School of Education, and another 2 assistants were members of the Rutgers Physics and Astronomy Education Group. Only 1 of the 8 teaching assistants was a traditional physics graduate student. In contrast, the teaching assistants for 203–204 included a mixture of several profes-

TABLE I. Measurements of students' evaluative abilities for unit analysis (UA) and special-case analysis (SCA) on each of the six exams.

Exam	Class	Fraction: UA	Fraction: SCA	Quality: UA	Quality: SCA
1	193–194	NA	(1.00)	NA	1.17 ± 0.07
	203–204		(1.00)		0.81 ± 0.05
2	193–194	0.89	0.18	2.48 ± 0.04	1.25 ± 0.08
	203–204	0.01	0.05	3.00 ± 0.00	2.33 ± 0.19
3	193–194	NA	(1.00)	NA	2.26 ± 0.04
	203–204		(1.00)		0.90 ± 0.05
4	193–194	0.53	0.38	2.73 ± 0.06	2.23 ± 0.07
	203–204	0.01	0.09	3.00 ± 0.00	2.30 ± 0.13
5	193–194	NA	0.42	NA	2.34 ± 0.08
	203–204		0.05		2.33 ± 0.19
6	193–194	0.38	0.38	2.35 ± 0.09	2.25 ± 0.07
	203–204	0.01	0.07	3.00 ± 0.00	2.25 ± 0.13

sors and several physics graduate students (none of whom were in physics education). These are, at least superficially, very distinct populations. This uncontrolled factor, and the threats it poses to the study's internal validity, shall be discussed below.

VIII. GOAL 1: TEACHING EVALUATION STRATEGIES

The 193–194 and 203–204 courses both had six exams during the year. Each exam featured a combination of multiple-choice questions and open-response tasks, with roughly two-thirds of the test score based on multiple-choice performance and one-third based on the open-response tasks. A typical midterm exam (exams 1, 2, 4, 5) had between 10 and 15 multiple choice questions, and 3 open-response tasks. Final exams (exams 3 and 6) had roughly double the number of items for each type. An evaluation task was included as an open-response task on each exam for both courses. These tasks served as summative assessments to measure students' evaluative abilities. Student responses to the evaluation tasks on each exam were photocopied and scored using our rubrics. Although we photocopied and scored each student's work from 193–194, the large number of students in 203–204 made it impossible to do the same for them. We therefore randomly selected 150 students from the 203–204 course whose responses to the evaluation tasks on exams would be photocopied and scored. A check was performed to determine whether this sample was representative of the entire class. A Kruskal-Wallis test indicated that the grades for the 203–204 sample were not significantly different from those of the remainder of the class.

The evaluation tasks on exams 1 and 3 were external special-case analysis tasks. These allowed us to measure how well the students could use this strategy when explicitly asked to. The evaluation tasks on exams 2, 4, 6 were critical thinking tasks. These allowed us to see what fraction of the students recognized that special-case and/or unit analysis could be used to construct such arguments, as would be demonstrated by their spontaneous use of these strategies. For those students that did attempt to use these strategies, the rubrics were used to measure the quality of their use. Each of these tasks involved topics which the 193–194 students had had both special-case and unit analysis tasks on (e.g., dc circuits). The evaluation task on exam 5 was a conceptual counterargument task, which allowed us to see what fraction of students tried to use special-case analysis, and how well they used it. For a list of the exam evaluation tasks, see Warren [37]. Table I lists the results from these tasks. The reported numbers for the quality of each strategy's use on exams 2, 4, 5, 6 are the average rubric scores of those students who decided to try using that particular strategy on the critical thinking tasks included in those exams. The values in columns 3 and 4 state the fractions of each class using each strategy, and were determined by finding the fraction of the class that had a score of at least 1 according to our rubrics. In column 4, a value of (1.00) indicates that students were explicitly instructed to use a particular strategy on an evaluation task. Values of NA in columns 3 and 5 indicate that no evaluation tasks relating to UA were included. All differences in fractional use between the two classes are statistically significant at the 0.01-level according to a χ^2-test. Note that there is an NA in column 4 for exams 1 and 3 because the tasks given on those exams explicitly told students to do a special-case analysis.

There were typically 8–12 students in our sample from 203–204 who tried using special-case analysis on the evaluation tasks from those exams, and their average quality of use according to our rubric was ~ 2.3 for each exam. No more than 1 student in our sample from 203–204 ever used unit analysis on the critical thinking tasks.

There are a few points to be made here. First, the results indicate that we were successful at helping the 193–194 students understand when, why, and how to use both strategies. They significantly outperformed the 203–204 students in the frequency of use for each evaluation strategy, and eventually demonstrated high quality use of each strategy as measured by the rubrics. The increases in special-case analysis quality from exam 1 to exam 3, and in the fraction using it from exam 2 to exam 4, are in part attributable to the fact that the

ratio of types of tasks given in recitation and homework assignments was altered after exam 2. Up until then, roughly half of the evaluation tasks included in these assignments focused on unit analysis while the other half dealt with special-case analysis. After seeing the results from exam 2, though, it was realized that students were having much more success with unit analysis, perhaps because it is a much simpler strategy than special-case analysis. Thereafter, the majority of evaluation tasks in recitations and homework focused on special-case analysis, and a substantial fraction (~40%) of the students are observed to spontaneously employ SCA on the exam evaluation activities.

It should be clarified that on exams 4 and 6, the set of 193–194 students who used unit analysis and the set of those that used special-case analysis were not identical even though the fractional usage was similar for both strategies. The overlap between these two sets (defined as the ratio of their intersection to their union) was 0.45 for exam 4 and 0.53 for exam 6.

It is interesting that special-case analysis was used at all among the 203–204 students on the critical thinking tasks on exams 2, 4, 5, 6, as a group of roughly 8–12 students spontaneously employed SCA with a high level of quality (average rubric scores of roughly 2.3). Apparently some students may enter our physics courses with a well-developed understanding of special-case analysis, perhaps due to prior physics courses. It is also worth noting that 203–204 students who tried using special-case analysis for the open-response exam problem typically scored very well on the multiple-choice questions, with 38% of them earning perfect scores.

IX. GOAL 2: USING AND VALUING EVALUATION STRATEGIES

As a means to assess student valuation and use of evaluation strategies, anonymous surveys were administered at the last recitation of the spring term in the Physics 194 course. Included were several questions asking students to rate how frequently they had used each strategy to evaluate their own work outside of class, and also to rate how strongly certain factors inhibited their use of each strategy. These survey questions are shown in Fig. 2, with the results in Table II. There were 158 respondents to the survey out of the 200 students in the course, giving a response rate of 79%. There may be a selection bias among the respondents, as the students knew that one recitation grade would be dropped from the final grade, and that the last recitation would be a review session instead of a normal recitation. For that reason, upper and lower bounds are listed on the average scores in Table II, where the bounds are determined by assuming all missing respondents would have given either the highest or lowest possible response for the questions.

To conduct a Cronbach alpha test for internal consistency, the scores to questions 3(a)–3(d) and 6(a)–6(d) were made negative, accounting for the fact that these inhibitors are likely to reduce the usage and valuation of evaluation strategies by students (i.e., all scores were coded in the same conceptual direction). The calculated alpha value is $\alpha = 0.800$, a strong result giving confidence that the items are consistent and can therefore be used as a basis for interpretation about student usage and valuation of evaluation strategies.

1. How much have you used special-case analysis on your own to learn/do physics?
1=Not at all 2=Not Much 3=Somewhat 4=Very 5=Extremely

2. If you were a physics major, and really interested in learning physics, how useful do you think special-case analysis would be for your learning?
1=Not at all 2=Not Much 3=Somewhat 4=Very 5=Extremely

3. Rate, on a scale from 0-10, how much each of the listed factors affected your desire to do special-case analysis (10 means the factor really made you *not* want to do special-case analysis ; 0 means the factor did not matter)
a) ___Time constraints (due to other classwork, jobs, etc.)
b) ___Motivation to learn physics (or lack thereof)
c) ___Confusion about how to do special-case analysis (if you weren't confused, put 0)
d) ___Confusion about the purpose of a special-case analysis (if you weren't confused, put 0)

4. How much have you used dimensional analysis on your own to learn/do physics?
1=Not at all 2=Not Much 3=Somewhat 4=Very 5=Extremely

5. If you were a physics major, and really interested in learning physics, how useful do you think dimensional analysis would be for your learning?
1=Not at all 2=Not Much 3=Somewhat 4=Very 5=Extremely

6. Rate, on a scale from 0-10, how much each of the listed factors affected your desire to do dimensional analysis (10 means the factor really made you *not* want to do dimensional analysis ; 0 means the factor did not matter)
a) ___Time constraints (due to other classwork, jobs, etc.)
b) ___Motivation to learn physics (or lack thereof)
c) ___Confusion about how to do dimensional analysis (if you weren't confused, put 0)
d) ___Confusion about the purpose of a dimensional analysis (if you weren't confused, put 0)

FIG. 2. End-of-year survey questions regarding the students' self-reported use of each evaluation strategy.

One may safely assume that there were very few students who came into the 193–194 course already knowing, valuing, and using either evaluation strategy. This assumption is supported by the small number of students from 203–204 who used either strategy on the tasks in exams 2, 4, 5, and 6 (see above). Given that assumption, the results suggest a moderate degree of success in teaching students to incorporate these strategies into their personal learning behavior. Indeed, responses to questions 1 and 4 were probably artificially lowered due to the fact that we only included evaluation tasks relating to half of the topics during the year. If evaluation tasks had been given on all topics, it seems likely that students would have used the evaluation strategies more frequently. It is worth pointing out that the evidence here has limited inferential strength because of its indirectness. Further studies with more direct means of assessing whether students learn when and why to employ evaluation strategies would be useful.

TABLE II. Results from end-of-year survey questions.

Question number	Average response	Lower bound	Upper bound
1	2.2/5	1.9	2.8
2	3.8/5	3.2	4.1
3a	7.0/10	5.5	7.6
3b	5.1/10	4.0	6.1
3c	2.4/10	1.9	4.0
3d	2.2/10	1.7	3.8
4	2.9/5	2.5	3.3
5	3.9/5	3.3	4.1
6a	5.9/10	4.7	6.7
6b	4.7/10	3.7	5.8
6c	1.7/10	1.3	3.4
6d	1.6/10	1.3	3.4

The relatively low response scores to questions 3c, 3d, 6c, and 6d are consistent with the results for goal one of our study, indicating we were successful at helping students understand when, why, and how to use each strategy. The fact that the scores for 3c and 3d are higher than 6c and 6d appears reasonable because special-case analysis is certainly a much more complicated and multifaceted strategy that unit analysis. This disparity in complexity probably also explains why the scores to 3a and 3b were higher than 6a and 6b.

X. GOAL 3: ENHANCING PROBLEM-SOLVING PERFORMANCE BY TEACHING EVALUATION

Several conventional multiple-choice questions were common to each of the 193–194 and 203–204 lecture exams. There were six exams during the year, three per semester. The midterm exams (exams 1, 2, 4, and 5) were 80 min, while the final exams (exams 3 and 6) were 3 h. The exam questions shared by the two classes were all designed by the instructor of the 203–204 course (Van Heuvelen). Some of these shared multiple-choice questions were on topics which the 193–194 students had had evaluation tasks on, such as work-energy and dc circuits. This set of multiple-choice questions will be called E-questions. The remainder of the shared multiple-choice exam questions covered topics that no one had had evaluation tasks on, such as momentum and fluid mechanics. These will be called NE-questions. The E- and NE-questions are listed in Appendix B of Warren [37] and are assumed to provide accurate measures of student problem-solving performance [38]. This assumption mitigates the strength of the results reported below, and more robust measures of problem-solving performance would be useful in future work. Roughly 80% of these questions were numerical, and the remaining 20% of questions were either conceptual or symbolic. Some of the conceptual questions were directly taken or modified from standardized assessment tools such as FCI [39] and CSEM [40]. All of the questions had been used by Van Heuvelen in previous years.

The first midterm exam had 2 NE-questions, and no E-questions, as there had not been any evaluation tasks used to this point in the 193–194 course. There were 2 E-questions and 2 NE-questions on midterm exams 2, 4, and 5. Exam 3 had 4 NE questions and 3 E-questions, and Exam 6 had 5 NE-questions and 4 E-questions.

Given this design, we may make two predictions based on the view of evaluation strategies as tools of self-regulated learning. First, because of the known population bias between the 193–194 and 203–204 students, the 203–204 students are expected to do better on the NE-questions. This prediction assumes that whatever benefits the evaluation tasks may have for the 193–194 students' E-question performance will not be transferable to the topics covered by NE-questions.

A second prediction is that the performance of the 193–194 students should be relatively better on E-questions than on NE-questions if the evaluation tasks succeed in benefiting student understanding of the topics covered by the tasks. Moreover, this boost of E-question performance should vary as students become more adept at the use of evaluation strategies. Therefore, it is predicted that student performance on the open-response evaluation tasks included on each exam will directly correlate with the relative performance of 193–194 students on E-questions.

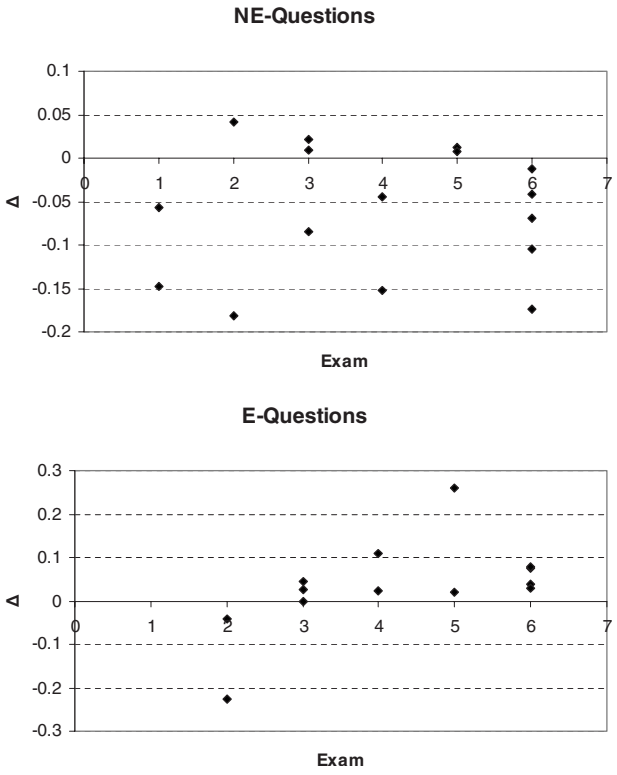


FIG. 3. Plots of the normalized difference between each condition on multiple-choice exam questions.

The relative performances of the 193–194 and 203–204 students on E- and NE-questions are compared in Fig. 3 and Table III. In Fig. 3, the normalized difference (Δ) between the two classes for each E- and NE-question is computed as

$$\Delta = \frac{C_{\text{exp}} - C_{control}}{C_{\text{exp}} + C_{control}},$$

where C_{exp} and $C_{control}$ denote the percentage of students in each condition who correctly answered the problem. A positive value therefore indicates that the experiment group (193–194) outperformed the control group (203–204) on a particular problem, and vice-versa for a negative value. The magnitude gives a measure of how much of a difference there was between groups, and it exaggerates differences on problems where the sum $C_{\text{exp}}+C_{control}$ is small. That is, it rewards improvements made on 'difficult' problems more than improvements made on "easy" problems.

The NE-question results show that the 203–204 students often did significantly better on problems relating to topics for which both classes had very similar learning environments, with the differences on 7 NE-questions being significant at the 0.01-level, and another 2 at the 0.05-level, all with the 203–204 sample doing better than the 193–194 sample. This observation was anticipated due to the known selection bias in this study. Given this bias, it would have been an achievement simply to bring the 193–194 students to a com-

TABLE III. Cross tabulation of performance on multiple choice (MC) exam problems, and the exact significance (2-tailed) for chi-squared tests of independence between each class. *=significant at 0.05 level, **=significant at 0.01 level.

Exam Problem	Class	% Correct	p-value	Exam Problem	Class	% Correct	p-value
Ex.1, NE 1	193–194	71.1	0.026*	**Ex.4, NE 1**	193–194	88.5	0.001**
	203–204	79.7			203–204	96.6	
Ex.1, NE 2	193–194	48.7	<0.001**	**Ex.4, NE 2**	193–194	41.1	0.001**
	203–204	65.5			203–204	55.9	
Ex.2, E 1	193–194	25.8	0.001**	**Ex.5, E 1**	193–194	68.6	<0.001**
	203–204	40.9			203–204	40.3	
Ex.2, E 2	193–194	81.3	0.049*	**Ex.5, E 2**	193–194	95.7	0.081
	203–204	88.0			203–204	91.5	
Ex.2, NE 1	193–194	53.8	0.361	**Ex.5, NE 1**	193–194	89.4	0.496
	203–204	49.4			203–204	87.0	
Ex.2, NE 2	193–194	52.2	<0.001**	**Ex.5, NE 2**	193–194	93.1	0.622
	203–204	75.4			203–204	91.8	
Ex.3, E 1	193–194	83.1	1.000	**Ex.6, E 1**	193–194	91.3	<0.001**
	203–204	83.2			203–204	78.7	
Ex.3, E 2	193–194	80.3	0.330	**Ex.6, E 2**	193–194	69.9	0.244
	203–204	76.4			203–204	64.5	
Ex.3, E 3	193–194	78.1	0.124	**Ex.6, E 3**	193–194	78.7	0.268
	203–204	71.2			203–204	74.0	
Ex.3, NE 1	193–194	80.9	0.380	**Ex.6, E 4**	193–194	54.6	0.072
	203–204	77.4			203–204	46.3	
Ex.3, NE 2	193–194	97.2	0.368	**Ex.6, NE 1**	193–194	60.1	0.050*
	203–204	95.4			203–204	69.0	
Ex.3, NE 3	193–194	69.1	0.001**	**Ex.6, NE 2**	193–194	50.3	<0.001**
	203–204	82.0			203–204	71.5	
Ex.3, NE 4	193–194	45.5	0.038*	**Ex.6, NE 3**	193–194	62.3	0.768
	203–204	55.1			203–204	64.0	
Ex.4, E 1	193–194	96.4	0.062	**Ex.6, NE 4**	193–194	63.9	<0.001**
	203–204	92.2			203–204	78.9	
Ex.4, E 2	193–194	91.1	<0.001**	**Ex.6, NE 5**	193–194	70.5	0.146
	203–204	73.1			203–204	76.5	

parable level of performance on the E-questions. In fact, the results show that by exam 3, the performance of the 193–194 students on E-questions was not only comparable to, but was better than that of the 203–204 students (although only three of the positive differences on E-questions were significant at the 0.01-level).

The favored hypothesis to explain these data is that the improved relative performance of the 193–194 students on E-questions was due to the use of our evaluation tasks. However, the strength of this hypothesis may be mitigated by the presence of uncontrolled factors in the study. Although students were not randomly assigned into the conditions, it is difficult to see how this may account for the observed differences on NE- and E-questions. Differences between the teaching populations for the two classes also fail to provide a reasonable mechanism for producing the differences on E-questions and NE-questions simultaneously. If there were any sort of "good teacher effect" it would be likely to affect the results for both E- and NE-questions.

Another threat to internal validity stems from the fact that the lecturers were not blind to the study, and may have unintentionally skewed the results. This potential experimenter bias can be argued against because of the fact that lectures for 193–194 and 203–204 were designed to be as similar as possible, and it is not at all clear how stylistic differences could cause such preferential performance differences between E- and NE-questions in any consistent fashion. All topics, whether those tested by E- or NE-questions, were covered in very similar fashions during lectures, recitations, and laboratories. Also, the fact that these performance differences developed and persisted through two different lecturers for 193–194 suggests that differences in lecture style were probably not a significant causal factor.

Another alternative hypothesis is that the results are due to time-on-topic differences in the recitation and homework assignments. Although recitation and homework assignments were designed to minimize such differences, no actual measurements were made. It is possible that the evaluation tasks

simply took longer for students to complete, and that their performance on E-questions was due not to the format of the activity but simply that they spent more time thinking about the concepts involved in the activity. However, anecdotal evidence from teaching assistants who helped students with their recitation and homework assignments suggests that students did not take an inordinate amount of time to complete the evaluation tasks. Also, there was no indication from student comments made to the teaching assistants that they felt the evaluation tasks took much longer than other recitation and homework problems.

While the results above indicate that the use of evaluation tasks benefited student problem-solving performance, we can further test this hypothesis by looking for a concrete association between the strength of student's evaluation abilities and their problem-solving performance. We construct two *ad hoc* measures to reflect students' apparent understanding of when, why, and how to use each evaluation strategy for a certain topic. The measures are

$$SCA_{relative} = FS_{\exp} \times QS_{\exp} - FS_{control} \times QS_{control},$$

$$UA_{relative} = FU_{\exp} \times QU_{\exp} - FU_{control} \times QU_{control},$$

where *SCA* stands for "special-case analysis," *UA* stands for "unit analysis," *FS* is the fraction of the class (experimental or control) which used special-case analysis and *FU* is the fraction which used unit analysis on the evaluation tasks included on exams 2, 4, 5, and 6. $UA_{relative}$ is not applicable for exam 5 due to the task format (conceptual counterargument). Also, *QS* is the average quality of the class' special-case analyses, and *QU* is the average quality of the class' unit analyses for these tasks. Each of these quantities corresponds to the appropriate values in Table I.

The fractions *FS* and *FU* serve as indicators of how well students understand *when* and *why* to use each strategy. If students do not understand the purpose of an evaluation strategy they are not likely to spontaneously employ it on the critical thinking or conceptual counterargument tasks without specific prompting. The quantities *QS* and *QU* are given by the evaluation ability rubrics, and measure the students' understanding of *how* to use each strategy. By taking the products $FS \times QS$ and $FU \times QU$, we get a pair of numbers between 0 and 3 which are taken to be indicative of each class' overall understanding of each evaluation strategy.

To measure relative student performance on E- and NE-questions, normalized differences in class performance for each problem category are averaged on each exam,

$$E_{relative} = \frac{1}{N_{E-problems}} \sum_{E-problems} \Delta_{E-problem},$$

$$NE_{relative} = \frac{1}{N_{NE-problems}} \sum_{NE-problems} \Delta_{NE-problem},$$

where Δ represents the normalized difference between the two classes' performance (as plotted in Fig. 3), and *N* denotes the number of problems of a certain type (either E- or NE-questions) on an exam.

TABLE IV. Measures of relative overall class performance for exams 1 through 6. The definitions of each measure are described in the text.

Exam	$SCA_{relative}$	$UA_{relative}$	$E_{relative}$	$NE_{relative}$
1	NA	NA	NA	−.102
2	0.179	2.047	−.045	−.070
3	NA	NA	0.024	−.037
4	0.645	1.272	0.066	−.098
5	0.851	NA	0.141	0.010
6	0.683	0.893	0.057	−.080

Table IV lists the values for these four measures of relative class performance on each exam. To determine whether special-case analysis and unit analysis performance related to performance on the exam questions, a Pearson correlation analysis of these data was performed. The results indicate that $E_{relative}$ is significantly positively correlated with $SCA_{relative}$ ($r=.996$, $p=.004$) and negatively correlated with $UA_{relative}$ ($r=-.921$, $p=.255$) while $NE_{relative}$ is not significantly correlated with both $SCA_{relative}$ ($r=.447$, $p=.553$) and $UA_{relative}$ ($r=.528$, $p=.646$).

The most parsimonious account for these results entails three hypotheses. One is that giving a greater proportion of evaluation tasks on special-case analysis after exam 2 (as discussed above) helped students to improve their use of that strategy while at the same time reducing their use of unit analysis. The fact that we manipulated these two independent factors in this fashion could thereby explain why the fractional usage of unit analysis steadily declined from exams 2 to 6 (see Table II).

A second hypothesis is that students' use of special-case analysis (in recitation, homework, and in their personal learning behavior) significantly benefited their problem-solving ability. The relative performance of the 193–194 students on E-questions correlated very strongly with their use of special-case analysis on the exam evaluation tasks. Note that because we are looking at relative differences in performance between the experiment and control groups, we can safely rule out the possibility that this correlation is due to the "easiness" of the subject matter or any other such factor.

The third hypothesis is that the use of unit analysis (in recitations, homework, and personal learning behavior) did not benefit students' problem-solving ability as much as special-case analysis. It would appear that the 193–194 students did not appreciate the greater utility of special-case analysis, though, since on the end-of-year survey they reported unit analysis as being just as valuable for learning physics as special-case analysis (see Table II). It therefore seems that the students had some limitations in their ability to self-assess the utility of special-case analysis and unit analysis for their learning. Also, the fact that $NE_{relative}$ is uncorrelated with both $UA_{relative}$ and $SCA_{relative}$ indicates that there is no transfer in the benefits of either strategy to topics not covered by the evaluation tasks from recitation and homework assignments.

TABLE V. Interpretive devices listed by class. Reproduced from Sherin [41] (Chapter 4, Fig. 2).

Narrative Class	Static Class
Changing Parameters	*Specific Moment*
Physical Change	*Generic Moment*
Changing Situation	*Steady State*
	Static Forces
	Conservation
	Accounting
Special Case	
Restricted Value	
Specific Value	
Limiting Case	
Relative Values	

XI. DISCUSSION

While the strategy of special-case analysis is rather complex, and took more time and effort for students to learn, it appears to provide significant benefits to performance on multiple-choice exam questions. In contrast, the simpler strategy of unit analysis is learned very quickly, but apparently provided little or no benefits on exam questions. Here we shall discuss some possible reasons for these results.

In his doctoral dissertation, Sherin [41] presented and discussed a categorization of interpretive strategies (which he calls "interpretive devices") used by students to give meaning to physics equations while working on back-of-chapter homework problems. His classification of these strategies is reproduced in Table V. There are three classes of devices; narrative, static, and special case.

Each of these devices plays a role in what we call special-case analysis (this article uses the term "special-case" in a much broader sense than Sherin). One may conduct many different special-case analyses of an equation, using any one of these interpretive devices as the specific means. For example, one may choose to analyze a case where some parameter is changed in the problem, or where a quantity is taken to some limiting value. By engaging students in special-case analyses through the use of our tasks, students are given the opportunity and feedback to practice the use of these interpretive devices.

Sherin argues that interpretive devices function as sense-making tools which build meaning around an equation by relating it to other pieces of knowledge. The field of semiotics studies exactly how humans make meaning using resources such as systems of words, images, actions, and symbols, and has identified two aspects of meaning called typological meaning (meaning by kind) and topological meaning (meaning by degree) [42]. Typological meaning is established by distinguishing categories of objects, relations, and processes. For example, an equation gains typological meanings by identifying categories of physical situations in which it is applicable (e.g., we can use $a=v^2/r$ for circular motion), the types of physical idealizations it assumes (e.g., we will assume v and r are constant), and by categorizing the quantities in the equation (e.g., does v correspond to voltage, or velocity? Does r correspond to the radius of the circle, or the size of the object?).

Topological meaning is created by examining changes by some degree. For example, an equation gains topological meanings by being relatable to other physical situations within the category of applicable situations (e.g., if r had been greater, how would a change?), or with situations which lie on the borderline of applicable situations (e.g., if we let $r=\infty$, what happens to a?). Also, an equation gains topological meaning by examining gradual deviations from the idealizations used by the equation (e.g., what would happen if v was increasing?). Typological and topological meaning for physics equations may be developed by a variety of activities [43].

Typological and topological meanings are not distinct, but necessarily relate to one another. For our example of centripetal acceleration, by taking r to infinity we enter a new category of physical situations, as the motion now has constant velocity. Therefore, this aspect of topological meaning construction develops typological meaning by highlighting the relation between the two categories of constant velocity motion and uniform circular motion.

So it is possible that the use of special-case analysis tasks, inasmuch as they compel students to use interpretive devices, aid student understanding of equations and consequently benefit their performance on exam questions. Unit analysis, on the other hand, makes no apparent use of interpretive devices and therefore seems incapable of constructing topological meaning. Unit analysis may help to construct some typological meaning, as it compels students to figure out which physical property each specific quantity corresponds to, but in general would seem to be weaker than special-case analysis at developing meaning for equations. This does not mean that unit analysis is a worthless strategy, as it certainly does serve the important function of testing for self-consistency in an equation. It may be that unit analysis could have a stronger impact if taught in a different fashion which places more emphasis on the conceptual aspects associated with it.

XII. CONCLUSION

Evaluation is a well-recognized part of learning, yet its importance is often not reflected in our introductory physics courses. This study has developed tasks and rubrics designed to help students become evaluators by engaging them in formative assessment activities. Results indicated that the use of evaluation activities, particularly those focusing on special-case analysis, help generate a significant increase in performance on multiple-choice exam questions. One possible explanation for this is that special-case analysis tasks engage students in the use of interpretive devices to construct typological and topological meaning for equations. It should be noted that increases in performance were only observed on E-questions, indicating a lack of transfer to topics not covered by the evaluation activities. Future studies designed to replicate and expand upon these results are necessary in order to strengthen the external validity of our results. Also, the

potential epistemological benefits of using evaluation tasks should be investigated, as well as the interactions with student motivation and self-efficacy.

ACKNOWLEDGMENTS

This study was conducted while the author was a member of the Rutgers Physics & Astronomy Education Group. I would like to thank Alan Van Heuvelen, Eugenia Etkina, Sahana Murthy, Michael Gentile, David Brookes, and David Rosengrant for valuable discussions and help with the execution of this project. Also, the feedback of three anonymous referees was helpful in improving the quality of this paper. This work was partially supported by the National Science Foundation (Grant No. DUE-0241078).

APPENDIX: EVALUATION ACTIVITIES

See separate auxiliary material for the evaluation tasks used in the 193–194 recitation and homework assignments, the evaluation rubrics developed and used, the exam evaluation tasks, and the E- and NE-problems.

[1] B. S. Bloom, *A Taxonomy of Educational Objectives, Handbook I: The Cognitive Domain* (David McKay Co. Inc., New York, 1956).

[2] *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*, edited by L. W. Anderson and D. R. Kraftwohl (Longman, New York, 2001).

[3] A. E. Lawson, How do humans acquire knowledge? And what does that imply about the nature of knowledge? Sci. Educ. **9**, 577 (2000).

[4] A. E. Lawson, The generality of hypothetico-deductive reasoning: Making scientific reasoning explicit, Am. Biol. Teach. **62**, 482 (2000).

[5] A. R. Warren, Evaluation Strategies: Teaching Students to Assess Coherence & Consistency, Phys. Teach. **47**, 466 (2009).

[6] A. Buffler, S. Allie, F. Lubben, and B. Campbell, The development of first year physics students' ideas about measurement in terms of point and set paradigms, Int. J. Sci. Educ. **23**, 1137 (2001).

[7] R. F. Lippmann, Ph.D. Dissertation, University of Maryland, 2003.

[8] S. P. Marshall, *Schemas in Problem-Solving* (Cambridge University Press, New York, 1995).

[9] W. J. Leonard, W. J. Gerace, and R. J. Dufresne, Analysis-based problem solving: Making analysis and reasoning the focus of physics instruction, Sci. Teach. **20**, 387 (2002).

[10] F. Reif and J. I. Heller, Knowledge structures and problem solving in physics, Educ. Psychol. **17**, 102 (1982).

[11] 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).

[12] M. Sabella and E. F. Redish, Knowledge organization and activation in physics problem-solving, University of Maryland pre-print (2004).

[13] R. R. Hake, Interactive-Engagement Versus Traditional Methods: A Six-Thousand-Student Survey of Mechanics Test Data for Introductory Physics Courses, Am. J. Phys. **66**, Issue 1, 64 (1998).

[14] K. Cummings, J. Marx, R. Thornton, and D. Kuhl, Evaluating innovation in studio physics, Am. J. Phys. **67**, S38 (1999).

[15] L. C. McDermott and P. S. Schaffer, *Tutorials in Introductory Physics* (Prentice-Hall, Upper Saddle River, NJ, 2002).

[16] N. D. Finkelstein and S. J. Pollock, Replicating and understanding successful innovations: Implementing tutorials in introductory physics, Phys. Rev. ST Phys. Educ. Res. **1**, 010101 (2005).

[17] A. P. Fagen, C. H. Crouch, and E. Mazur, Peer Instruction: Results from a range of classrooms, Phys. Teach. **40**, 206 (2002).

[18] E. Mazur, *Peer Instruction: A User's Manual* (Prentice-Hall, Upper Saddle River, NJ, 1996).

[19] R. Warnakulasooriya, D. J. Palazzo, and D. E. Pritchard, in *Evidence of Problem-Solving Transfer in Web-Based Socratic Tutor*, Physics Education Research Conference Proceedings, edited by P. Heron, L. McCullough, and J. Marx (AIP Press, Melville, 2006).

[20] K. VanLehn, C. Lynch, K. Schulze, J. Shapiro, R. Shelby, L. Taylor, D. Treacy, A. Weinstein, and M. Wintersgill, The Andes Physics Tutoring System: Lessons Learned, Int. J. Artif. Intell. Educ. **15**, No. 3, 147 (2005).

[21] B. J. Zimmerman and M. Martinez-Pons, Development of a structured interview for assessing student use of self-regulated learning strategies, Am. Educ. Res. J. **23**, Issue 4, 614 (1986).

[22] N. Perry, Young children's self-regulated learning and contexts that support it, J. Educ. Psychol. **90**, 715 (1998).

[23] D. Hammer, More than misconceptions: Multiple perspectives on student knowledge and reasoning, and an appropriate role for education research, Am. J. Phys. **64**, 1316 (1996).

[24] American Association for the Advancement of Science, *Project 206, Benchmarks for Science Literacy* (Oxford University Press, New York, 1993).

[25] NSF Directorate for Education and Human Resources Review of Undergraduate Education, *Shaping the Future: New Expectations for Undergraduate Education in Science, Mathematics, Engineering, and Technology*. Recommendations may be seen at http://www.ehr.nsf.gov/egr/due/documents/review/96139/four.htm (1996).

[26] National Research Council, *National Science Education Standards* (National Academy Press, Washington, D.C., 1996).

[27] R. H. Ennis, in *Teaching Thinking Skills: Theory and Practice*, edited by J. B. Baron and R. J. Sternberg (Freeman, New York, 1987), pp. 9–26.

[28] P. M. King and K. S. Kitchener, *Developing Reflective Judgment: Understanding and Promoting Intellectual Growth and Critical Thinking in Adolescents and Adults* (Jossey-Bass, San Francisco, 1994).

[29] P. Black and D. Wiliam, *Inside the Black Box: Raising Standards through Classroom Assessment* (King's College, Lon-

don, 1998).
[30] R. Sadler, Instr. Sci. **18**, 119 (1989).
[31] S. Murthy, *Peer-Assessment of Homework Using Rubrics*, 2007 Physics Education Research Conference, edited by L. Hsu, C. Henderson, and L. McCullough (AIP, Melville, NY, 2007), p. 156–159.
[32] E. Yerushalmi, C. Singh, and B. S. Eylon, *Physics Learning in the Context of Scaffolded Diagnostic Tasks (I)*, 2007 Physics Education Research Conference, edited by L. Hsu, C. Henderson, and L. McCullough (AIP, Melville, NY, 2007), p. 27–30.
[33] C. Singh, E. Yerushalmi, and B. S. Eylon, *Physics Learning in the Context of Scaffolded Diagnostic Tasks (II): Preliminary Results*, 2007 Physics Education Research Conference, edited by L. Hsu, C. Henderson, L. McCullough (AIP, Melville, NY, 2007), p. 31–34.
[34] E. Etkina, A. Van Heuvelen, S. Brahmia, D. Brookes, M. Gentile, S. Murthy, D. Rosengrant, and A. Warren, Scientific abilities and their assessment, Phys. Rev. ST Phys. Educ. Res. **2**, 020103 (2006).
[35] J. Cohen, A coefficient of agreement for nominal scales, Educ. Psychol. Meas. **20**, 37 (1960).
[36] A. R. Warren, *The Role of Evaluative Abilities in Physics Learning*, 2004 Physics Education Research Conference Proceedings, edited by J. Marx, S. Franklin, and P. Heron (AIP, Melville, NY, 2005).
[37] A. R. Warren, Ph.D. Dissertation, Rutgers University, 2006.
[38] M. Scott, T. Stelzer, and G. Gladding, Evaluating multiple-choice exams in large introductory physics courses, Phys. Rev. ST Phys. Educ. Res. **2**, 020102 (2006).
[39] D. Hestenes, M. Wells, and G. Swackhamer, Force concept inventory, Phys. Teach. **30**, 141 (1992).
[40] D. P. Maloney, T. L. O'Kuma, C. J. Hieggelke, and A. V. Heuvelen, Surveying students' conceptual knowledge of electricity and magnetism, Am. J. Phys. **69**, S12 (2001).
[41] B. Sherin, Ph.D. Dissertation, University of California, Berkeley, 1996.
[42] J. L. Lemke, in *"Multiplying Meaning: Visual and Verbal Semiotics in Scientific Text," in Reading Science*, edited by J. R. Martin and R. Veel (Routledge, London, 1998).
[43] E. Etkina, A. R. Warren, and M. J. Gentile, The Role of Models in Physics Instruction, Phys. Teach. **44**, 34 (2006).