

Evaluation Strategies as a Means for Learning Physics

Aaron R. Warren

Department of Physics & Astronomy, Rutgers University

Piscataway, NJ, 08854, USA

The Rutgers PAER Group has been working to help students develop a set of scientific abilities¹, including the ability to evaluate information. 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. In this paper I present results of research to investigate the feasibility and impact of teaching 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 their problem-solving performance.

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 as the only way for them to identify and learn from their mistakes. I argue that this dependence has several negative effects on students, detracting from their ability to learn both the content and process of physics.

Evaluation Strategies

The ability to effectively evaluate information has long been recognized as an important cognitive process and educational objective^{2,3}. According to Anderson & Kraftwohl, "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 some standard. In physics, there are a 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. My research 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 I believe it is important for students to learn and use evaluation strategies, I will outline two of these strategies. Each of these strategies relies upon hypothetico-deductive reasoning⁴⁵, whereby the information being evaluated is used to create a hypothesis which is then tested. The logical sequence for this testing can be characterized as: *If (general hypothesis) and (auxiliary assumptions) then (expected result) and/but (compare actual result to expected result), therefore (conclusion)*. Below I describe the type of information each strategy is meant to evaluate, the criteria by which the information is judged, and the hypothetic-deductive process that is used to judge the information. I also briefly describe how the results of each strategy may be used to help correct mistakes that have been identified.

Unit Analysis

Unit analysis is used to evaluate equations to determine whether they are physically sensible. A physically sensible equation should have the same unit for each term. The hypothetico-deductive process is:

If the equation is physically coherent,

And I correctly remember the units for each quantity in the equation,

Then I expect the units for each term in the equation to be identical,

And/But the units for each term are/are not identical,

Therefore the equation is/is not physically self-consistent.

If the equation is found to be physically incoherent, 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.

a) If the equation is incorrect, the student should determine exactly how the equation fails the unit analysis. The goal here is to 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 correct the equation on dimensional grounds to make it physically coherent.

b or c) If the student incorrectly remembered some units or made a mistake in their analysis, then the unit analysis has produced a false result.

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. Each equation, model, or claim is meant to be true for some range of physical situations. Our strategy is to choose some specific situation which we have prior knowledge of, and which lies within or at the limits of this range of applicability. We then determine what the equation, model, or claim predicts for this situation, and compare that with our knowledge of what actually happens. Basically, the approach here is to do a thought experiment to determine whether the equation, model, or claim makes sense based on what we already know. The hypothetico-deductive process is:

If the equation/model/claim is consistent with my prior knowledge for a specific situation,

And my prior knowledge for that specific situation is correct,

Then the equation/model/claim should predict a result which matches my prior knowledge for that specific situation,

And/But the prediction does/does not match my prior knowledge,

Therefore the equation/model/claim is/is not consistent with my prior knowledge.

If the equation/model/claim is found to be inconsistent with prior knowledge, there are three possible reasons: (a) the equation/model/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.

a) If the equation/model/claim is incorrect, the student should determine exactly how the equation/model/claim fails the special-case analysis. The goal here is to figure out how the equation/model/claim needs to be modified in order to make it consistent with the prior knowledge.

b or c) If the student's prior knowledge is incorrect, then the special-case analysis has produced a false result. The risk of this occurring can be minimized by having students evaluate special-cases for which they are highly confident in their prior knowledge. If the student made a mistake in executing the analysis, then the obtained result is false.

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. For example, if a student drops an exponent while doing a unit analysis, the result of their unit analysis is likely to be either a false positive or false negative (depending on the exact circumstances). This means that one can never say with absolute confidence that a negative or positive result of an evaluation is in fact correct. Furthermore, it is possible for an equation to pass several evaluations yet still be incorrect. For example, the equation $p = (1/2)mv$ passes a unit analysis although it is mistaken. Therefore the evaluation strategies listed above provide a useful though imperfect means for judging the validity and soundness of theoretical models and equations.

It is also important to note that each evaluation strategy, in and of itself, is context-independent. Of course, a significant amount of context-specific knowledge is required to properly activate and execute each strategy, but the strategy itself remains the same despite the context. For example, to do a unit analysis I must understand the general logical motivation and method for the analysis, besides knowing the correct units for each quantity in the expression at hand. The hypothetico-deductive steps for each strategy listed above constitute schematic knowledge which requires context-specific procedural and declarative knowledge to be properly activated and used.

Finally, there are certainly many other evaluation strategies used in physics, such as error analysis to evaluate an experimental result. 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, I 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.

Evaluation and Learning

Evaluation strategies may be modeled as schemas⁸, which I call *evaluation meta-schemas* because they serve to regulate the development and modification of other schemata. An

evaluation meta-schema may be linked to an array of schemata responsible for context-specific activities, such as solving inclined-plane problems. When one of these context-specific schemata are activated, an evaluation meta-schema (i.e., an evaluation strategy) has some probability of being activated as well. Subsequent use of an evaluation meta-schema 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 schema. Upon such recognition, the student may correct her own mistake, consequently restructuring the associated schema, or the links between several schemata. In other words, I believe that evaluation meta-schemas are responsible for establishing *local* and *global coherence*⁹, and also for modifying the conditions under which a particular schema is activated.

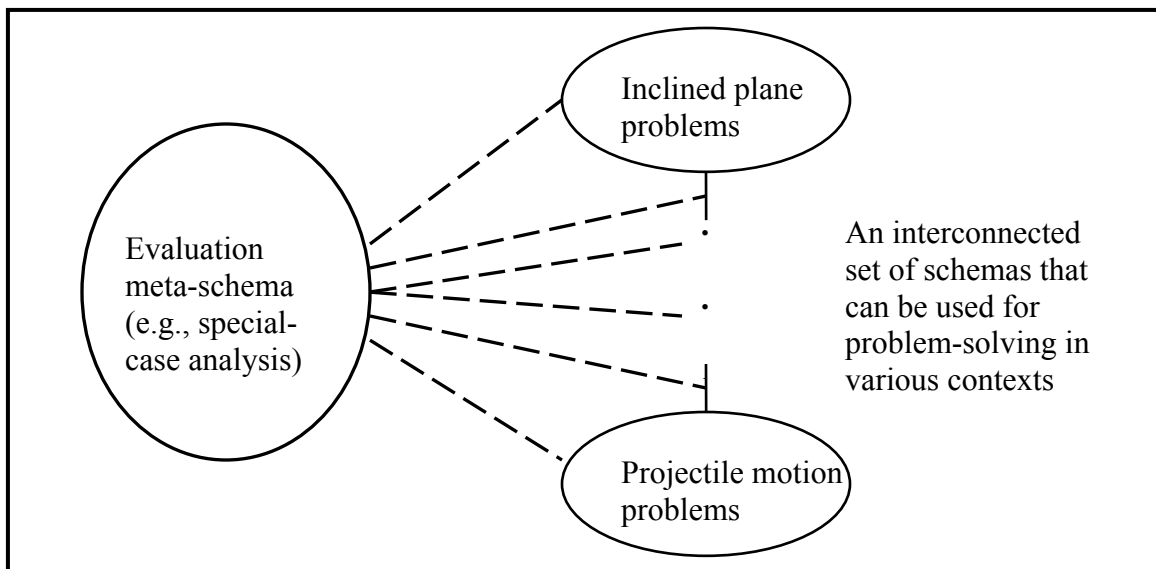
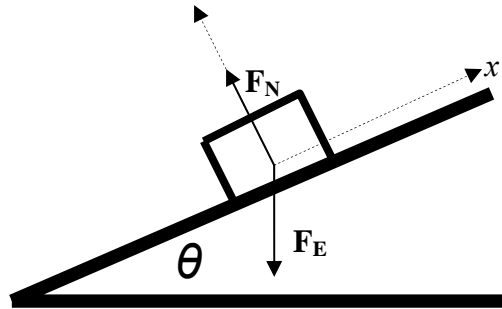


Figure 1. An evaluation strategy may function as a regulatory agent (called a meta-schema), capable of restructuring and reorganizing schemata related to various topics in order to establish coherence and consistency.

As a simple illustrative example, consider a student working on the homework problem shown in Figure 2. 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 $\theta = 0^\circ$, the student may realize that she has made a mistake. If $\theta = 0^\circ$, then we expect $F_N = Mg$ since the incline will be level, requiring the normal force to completely balance the gravitational force. However, plugging $\theta = 0^\circ$ into the student's solution gives $F_N = 0\text{N}$. 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, the 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. It is in this way that I believe the use of evaluation strategies serves a regulatory function in the organization of student knowledge.

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:

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

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

Therefore,

$$F_N = Mg \sin(\theta)$$

Figure 2. An illustrative example of a traditional homework problem and a possible student solution.

Without this evaluation meta-schema, a student's ability to revise and regulate her understanding is much more limited. When an instructor evaluates a student's work and identifies some mistake, the student may be unable to understand why and how her work was flawed because she lacks the evaluation meta-schema necessary for understanding why her mistake is, in fact, a mistake. Therefore, the student's schemas are less likely to be properly modified, meaning she is less likely to develop a robust understanding of physics.

If the sole means for evaluation lie outside our students, then they are forced into dependency upon some 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^{10 11}, such as Tutorials¹², Peer Instruction¹³, and intelligent tutoring systems^{14 15}. In each case, a student is given greater access to external evaluation, and hence has a greater opportunity to learn.

Although restructuring physics courses to enable more evaluative interaction does help, it can only help so much. 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 learn better on their own. In fact, Zimmerman and Martinez-Pons identified self-evaluation as a necessary component in their model of self-regulated learning¹⁶, and self-evaluation has been shown to promote self-regulated learning in young students¹⁷. Likewise, Hammer¹⁸ 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. This is a major goal of science education, as stated by many documents and papers^{19 20 21}. Authentic reasoning abilities rely upon an authentic epistemology, as someone who does not value consistency or understand the importance of uncertainty will probably never develop the reasoning abilities needed to deal with such subjects. However, if we can help students learn when, why, and how to evaluate their own work using strategies such as special-case analysis, then they may be more likely to recognize the value of establishing coherence and consistency among their knowledge, and develop at least some of the general reasoning abilities that are used in authentic science inquiries and practices.

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. This recognition of the possibility of errors is central to many aspects of critical thinking²² and reflective judgment²³. It seems reasonable to believe, then, that these abilities will be enhanced if we teach students to use evaluative strategies in physics.

Teaching & Assessing Evaluation

When teaching introductory physics, there are often several difficult constraints imposed upon the course. Among the most restricting of these constraints is the necessity for students to develop a working understanding of a wide range of subjects. Thus the available time which the class can spend on any single topic is often quite brief. Given this constraint, it is clear that any effort to explicitly teach evaluation to students must strive to have a maximal effect while demanding a minimal amount of instructional time.

Formative assessment activities may provide the best means for doing this. As defined by Black and Wiliam, formative assessment activities are "all those activities undertaken by teachers, and by their students in assessing themselves, which provide information to be used as feedback to modify the teaching and learning activities in which they are engaged."²⁴ After reviewing 580 articles and book chapters, they concluded that ongoing assessments by teachers, combined with appropriate feedback to students, can have powerful positive effects on student learning and achievement. In particular, learning gains produced by effective use of formative assessment are larger than those found for any other educational intervention (with effect sizes of 0.4 to 0.7).

But what, exactly, constitutes a formative assessment activity? Sadler²⁵ suggested three guiding principles, stated in the form of questions, that define formative assessment:

1. Where are you trying to go? (Identify and communicate the learning/performance goals)
2. Where are you now? (Assess, or help the student to self-assess, current levels of understanding.)

3. How can you get there? (Help the student with strategies and skills to reach the goal.)

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 rubric contains descriptions of different levels of performance, including the target level. A student or a group of students can use the rubric to self-assess her or their own work, or an instructor can use the rubric to assess students' responses and to provide feedback. In the context of teaching evaluative abilities, this means that the rubrics can be used to provide assessments of a student's ability to internally evaluate their own work with specific strategies. Additionally, written and verbal comments on student work may be given to personalize the feedback and address particular issues in a student's work.

Evaluation Tasks & Rubrics

During the 2002/3 and 2003/4 academic years, we developed, tested, and refined a library of tasks designed to help students learn when and how to use 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 does not understand the general method for completing the task, 2 = student understands how to perform the task in general, but his/her actual 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 [1]. Our final set of evaluation rubrics yielded an average inter-rater agreement level of 96%, ranging between 91% and 99%, and a Cohen Kappa of 0.947 ($p = .000$). The rubrics are available for download at <http://paer.rutgers.edu/ScientificAbilities/Rubrics/default.aspx>.

Study Design & Results

Preliminary research conducted during the 2003/4 academic year demonstrated significant correlations between students' abilities to use evaluative strategies and their problem-solving performance²⁶. During the 2004/5 academic year, I conducted a study to further test and investigate the effects of using evaluation tasks in an algebra-based physics course. After discussing the design of the study, I present results addressing each of 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 caused significant improvements in student problem-solving performance.

To accomplish these goals I collected and analyzed data from two courses at Rutgers University as part of a quasi-experimental control group study. The experiment group

consists of the 193/194 course, which is a year-long introductory algebra-based 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, which is also an introductory algebra-based course for life science majors. There were 459 students for the fall term and 418 students for the spring term. These two courses were generally run in parallel, although the 203/4 class covers slightly more material, and the students in this class typically have stronger math and science backgrounds than in the 193/4 course (as it is the more competitive course, held on a different university campus). This bias will actually serve to accentuate the results of our study, as discussed below.

The lectures for 203/4 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 myself. All lectures for 193 and 194 were based on Van Heuvelen's lecture notes. Lectures for all courses involved chalkboard/transparency-based presentations featuring experimental demonstrations and student engagement via peer discussions and student infrared response systems.

One premise of the study was the ability to provide very similar lecture environments for the two courses. In general, we believe we were successful in fulfilling this. However, some stylistic and personal differences in lecturing were unavoidable, and the effect of

these differences is not known. We will discuss this threat to the study's internal validity at the end of this paper.

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 203/4 were generally identical to those given in 193/4 except that some problems from the 203/4 assignments would be replaced by evaluation tasks. 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 took 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 203/4 course is required to cover more material, and has 55-minute recitations while the 193/4 course has 80-minute recitations. Again, we will address the threats posed by such factors later.

The evaluation tasks used in recitations and homework for 193/4 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 we did include tasks relating to work-energy, the First Law of Thermodynamics, and DC circuits (see Warren²⁷ 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 labs featured three types of experiments; investigative, testing, and application. Investigative experiments engage students in constructing a model for a novel phenomenon, testing experiments engage students in the design and execution of experiments to test a model, and application experiments engage students in the design and execution of experiments which use a model to determine some physical characteristic or produce a specific desired physical result. Each lab usually included two experiments, most of which were either testing or application experiments. It should be noted that although the 193/4 students were required to take the lab, the 203/4 students were not, as the labs constitute a distinct course (labeled 205/6). However, nearly all (~98%) of the 203/4 students were enrolled in 205/6 concurrently.

Another important factor in the study is the teachers themselves. For one thing, the teaching assistants in 193/4 are often not in a traditional physics program. In fact, 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 203/4 included a mixture of several professors and

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

Although this is not a true control group study, something that is nearly impossible to achieve in a classroom setting, it does provide a useful means for examining the feasibility and potential benefits of teaching evaluation strategies in introductory physics courses. Like any classroom study, this work was conducted in a very particular learning environment, with a particular population of students and teachers, thereby limiting the external validity of the results.

Goal 1: Teaching Evaluation Strategies

The 193/4 and 203/4 courses both had six exams during the year. Each exam featured a combination of multiple-choice problems and open-response problems. An evaluation task was included as an open-response problem 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 193/4, the large number of students in 203/4 made it impossible to do the same for them. We therefore randomly selected 150 students from the 203/4 course whose responses to the evaluation tasks on exams would be photocopied and scored.

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, we used our rubrics to measure the quality of their use. Each of these tasks involved topics which the 193/4 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 Warren²⁶. Table 1 lists the results from these tasks, and Figure 3 illustrates the data. The reported numbers for the quality of each strategy's use on exams 2, 4, 5, 6 are the average scores of those students who decided to try using that particular strategy on the critical thinking tasks included in those exams.

Exam	Class	Fraction: Unit Analysis	Fraction: Special-case Analysis	Quality: Unit Analysis	Quality: Special-case Analysis
1	193/4 203/4	NA	NA	NA	1.17 0.81
2	193/4 203/4	0.89 0.01	0.18 0.05	2.48 3.00	1.25 2.33
3	193/4 203/4	NA	NA	NA	2.26 0.90
4	193/4 203/4	0.53 0.01	0.38 0.09	2.73 3.00	2.23 2.30
5	193/4 203/4	NA	0.42 0.05	NA	2.34 2.33
6	193/4 203/4	0.38 0.01	0.38 0.07	2.35 3.00	2.25 2.25

Table 1. Measurements of students' evaluative abilities on each of the six exams.

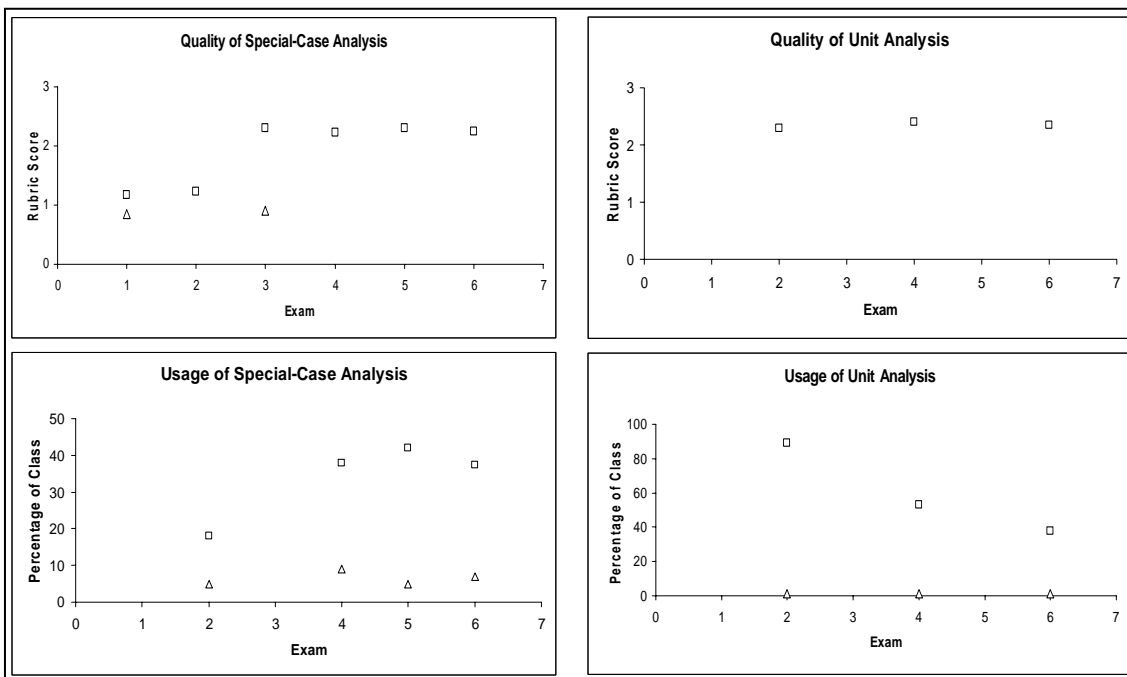


Figure 3. Plots of the data in Table 1. Triangles represent data from 203/4, squares represent data from 193/4.

Because there were so few students in 203/4 that used special-case or unit analysis on exams 2, 4, 5, 6, we have not plotted the quality of their use in Figure 2. There were typically 8-12 students in our sample from 203/4 who tried using special-case analysis on the evaluation tasks from those exams, and their average quality of use was ~2.4 for each exam. No more than 1 student in our sample from 203/4 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 193/4 students understand when, why, and how to use both strategies. They significantly outperformed the 203/4 students in both the frequency and quality of use for each evaluation strategy. The increase in special-case analysis abilities from exams between exams 1 and 3 is in part attributable to the fact that we altered the ratio of types of tasks given in recitation and homework assignments. Up until exam 2, 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, probably because it is a much simpler strategy than special-case analysis. So after exam 2, the majority of evaluation tasks in recitations and homework focused on special-case analysis. This emphasis is most likely responsible for the jump in performance on special-case analysis between exams 1 and 3, and also for the steady decline of unit analysis performance from exams 1 to 3 to 6.

It should be noted that on exams 4 and 6, the set of 193/4 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 .45 for exam 4 and .53 for exam 6.

It is interesting that special-case analysis was used at all among the 203/4 students on exams 2, 4, 5, 6. Apparently some students may enter our physics courses with a well-

developed understanding of special-case analysis, although these students may have had prior physics courses. It is also worth noting that 203/4 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.

Goal 2: Using & Valuing Evaluation Strategies

As a means to assess student valuation and usage of evaluation strategies, I administered anonymous surveys at the end of the spring term in the Physics 194 course. The surveys were given during the last recitation of the semester. Included were several questions asking the 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 Figure 4, and the results are shown in Table 2 and Figure 5. 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. Also, the last recitation was designed as a review session instead of being a normal recitation. Thus, the results here should be viewed as having some error associated with them due to this potential selection bias. For that reason, I list upper and lower bounds on the average scores in Table 2, 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)-(d) and 6(a)-(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.

I believe it can be safely assumed that there were very few students who came into the 193/4 course already knowing, valuing, and using either evaluation strategy. This assumption is supported by the small number of students from 203/4 who used either strategy on the tasks in exams 2, 4, 5, and 6 (see above). Also, anecdotal evidence from teaching assistants indicated that the 193/4 students were generally not aware of either evaluation strategy at the beginning of the year.

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)

Figure 4. End-of-year survey questions regarding the students' private use of each evaluation strategy.

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

Table 2. Results from end-of-year survey questions.

Given that assumption, our results show 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 we had given evaluation tasks relating to all topics, it seems likely that students would have used the evaluation strategies more frequently and for more topics than they actually did.

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 multi-faceted strategy than unit analysis. This disparity in complexity probably also explains why the scores to 3a and 3b were higher than 6a and 6b.

Goal 3: Enhancing Problem-Solving Performance by Teaching Evaluation

Several conventional multiple-choice problems were common to each of the 193/4 and 203/4 lecture exams. There were 6 exams during the year, 3 per semester. The mid-term exams (exams 1, 2, 4, and 5) were 80-minutes, while the final exams (exams 3 and 6) were 3 hours. The exam problems shared by the two classes were all designed by the instructor of the 203/4 course (Van Heuvelen). Some of these shared multiple-choice problems were on topics which the 193/4 students had had evaluation tasks on, such as work-energy and DC circuits. This set of multiple-choice problems will be called E-problems. The remainder of the shared multiple-choice exam problems covered topics that no one had had evaluation tasks on, such as momentum and fluid mechanics. These will be called NE-problems. The E- and NE-problems are listed in Warren²⁶. There were roughly 2 E-problems and 2 NE-problems on each mid-term exam (exams 1, 2, 4, and 5), and 4 E-problems and 4 NE-problems on the final exams (exams 3 and 6). The only major exception to this is exam 1, which did not have any E-problems.

Given this design, we may make several predictions based on the view of evaluation strategies as regulatory meta-schemas. First, because of the known population bias between the 193/4 and 203/4 students, the 203/4 students are expected to do better on the NE-problems. This prediction assumes that whatever benefits the evaluation tasks may have for the 193/4 students' E-problem performance will not be transferable to the topics covered by NE-problems. More specifically, it assumes that students will not spontaneously use the evaluation strategies in the context of topics that were not addressed by the evaluation tasks given in recitation and homework. This assumption is tested below.

A second prediction is that the performance of the 193/4 students should be relatively better on E-problems than on NE-problems if the evaluation tasks succeed in benefiting student understanding of the topics covered by the tasks. Moreover, this boost of E-problem performance should vary as the strength of the students' evaluation meta-schemas varies. Therefore, it is predicted that student performance on the evaluation tasks included on each exam will directly correlate with the relative performance of 193/4 students on E-problems.

The relative performance of the 193/4 and 203/4 students on E- and NE-problems is compared in Figure 5 and Table 3. In Figure 6, the normalized difference between the two classes for each E- and NE-problem is computed as:

$$\text{Normalized Difference} = (C_{\text{experiment}} - C_{\text{control}}) / (C_{\text{experiment}} + C_{\text{control}})$$

where $C_{\text{experiment}}$ and C_{control} denote the percentage of students in each group who correctly answered the problem. A positive value therefore indicates that the experiment group (193/4) outperformed the control group (203/4) on a particular problem, and vice-versa for a negative value.

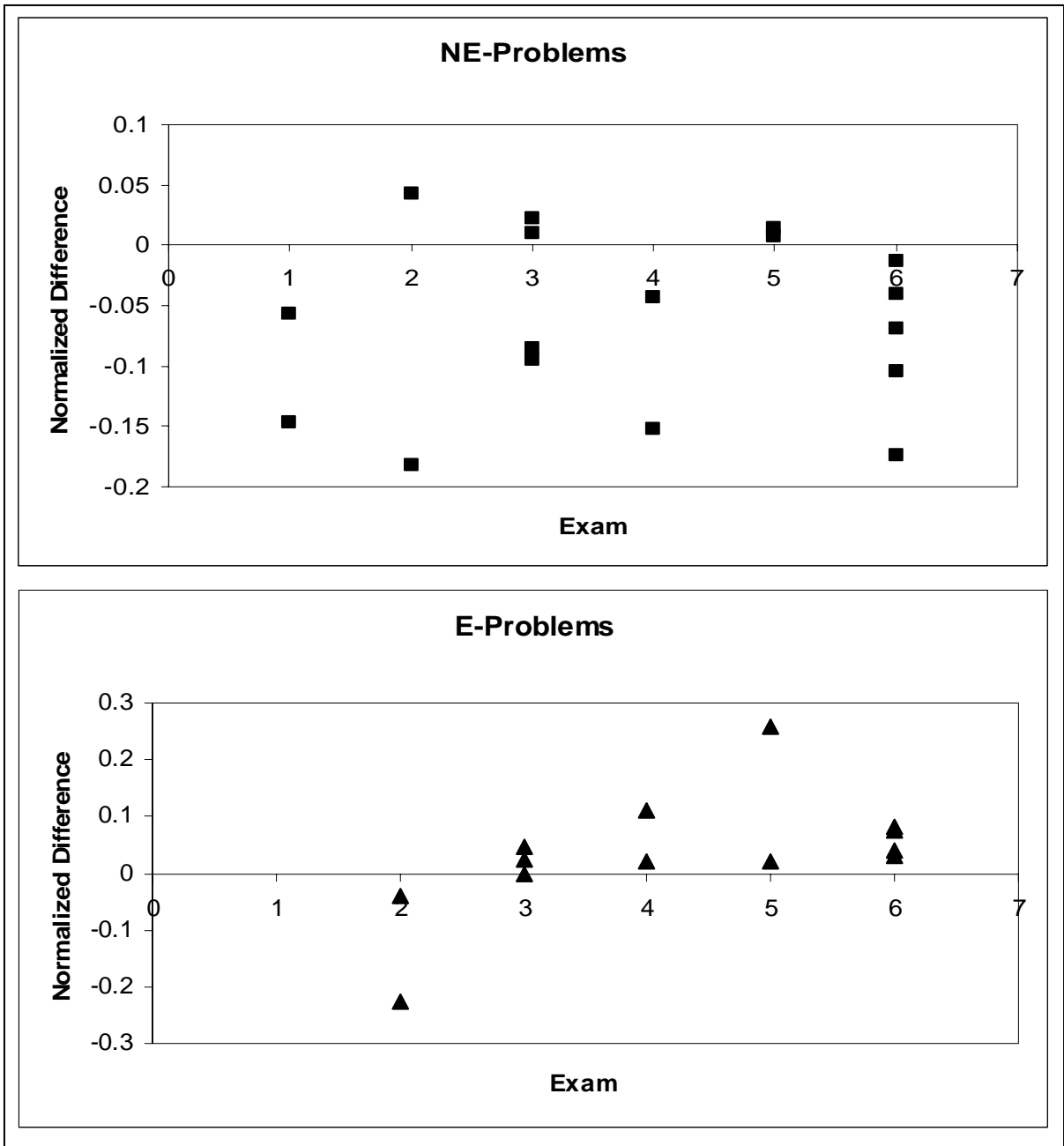


Figure 5. Plots of the normalized difference between each class on shared multiple-choice exam problems.

Exam Problem	Class	% Correct	<i>p</i>-value	Exam Problem	Class	% Correct	<i>p</i>-value
Ex.1, NE 1	193/4 203/4	71.1 79.7	.026*	Ex.4, NE 1	193/4 203/4	88.5 96.6	.001**
Ex.1, NE 2	193/4 203/4	48.7 65.5	.000**	Ex.4, NE 2	193/4 203/4	41.1 55.9	.001**
Ex.2, E 1	193/4 203/4	25.8 40.9	.001**	Ex.5, E 1	193/4 203/4	68.6 40.3	.000**
Ex.2, E 2	193/4 203/4	81.3 88.0	.049*	Ex.5, E 2	193/4 203/4	95.7 91.5	.081
Ex.2, NE 1	193/4 203/4	53.8 49.4	.361	Ex.5, NE 1	193/4 203/4	89.4 87.0	.496
Ex.2, NE 2	193/4 203/4	52.2 75.4	.000**	Ex.5, NE 2	193/4 203/4	93.1 91.8	.622
Ex.3, E 1	193/4 203/4	83.1 83.2	1.000	Ex.6, E 1	193/4 203/4	91.3 78.7	.000**
Ex.3, E 2	193/4 203/4	80.3 76.4	.330	Ex.6, E 2	193/4 203/4	69.9 64.5	.244
Ex.3, E 3	193/4 203/4	78.1 71.2	.124	Ex.6, E 3	193/4 203/4	78.7 74.0	.268
Ex.3, NE 1	193/4 203/4	80.9 77.4	.380	Ex.6, E 4	193/4 203/4	54.6 46.3	.072
Ex.3, NE 2	193/4 203/4	97.2 95.4	.368	Ex.6, NE 1	193/4 203/4	60.1 69.0	.050*
Ex.3, NE 3	193/4 203/4	69.1 82.0	.001**	Ex.6, NE 2	193/4 203/4	50.3 71.5	.000**
Ex.3, NE 4	193/4 203/4	45.5 55.1	.038*	Ex.6, NE 3	193/4 203/4	62.3 64.0	.768
Ex.4, E 1	193/4 203/4	96.4 92.2	.062	Ex.6, NE 4	193/4 203/4	63.9 78.9	.000**
Ex.4, E 2	193/4 203/4	91.1 73.1	.000**	Ex.6, NE 5	193/4 203/4	70.5 76.5	.146

Table 3. Crosstabulation of performance on MC-exam problems, and the exact significance (2-tailed) for chi-squared tests of independence between each class' performance on each question. * = significant at .05 level, ** = significant at .01 level.

The NE-problem results show that the 203/4 students often did significantly better on problems relating to topics for which both classes had very similar learning environments. 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 193/4 students to a comparable level of performance on the E-problems. In fact, the results show that by exam 3, the performance of the 193/4 students on E-problems was not only comparable to, but was better than that of the 203/4 students.

The favored hypothesis to explain these data is that the improved relative performance of the 193/4 students on E-problems was caused by the use of our evaluation tasks.

However, the strength of this hypothesis is mitigated by the presence of uncontrolled factors in the study. These uncontrolled factors each lend weight to a few alternative hypotheses that must be assessed as threats to the internal validity of this study.

For one thing, this study clearly did not randomly allocate students between the two courses. While we expected this selection bias to give the 203/4 class a higher relative performance on NE-problems, perhaps it was also responsible for the higher relative performance of 193/4 students on E-problems. The difficulty with this hypothesis is the

lack of any clear mechanism for such a result on the basis of selection differences alone, and for that reason I feel this hypothesis should not receive much weight in consideration. Differences between the teaching populations for the two classes also fail to provide a reasonable mechanism for producing the differences on E-problems and NE-problems simultaneously. If there were any sort of “good teacher effect” it would be likely to affect the results for both E- and NE-problems.

Another threat to internal validity stems from the fact that the lecturers were obviously 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 193/4 and 203/4 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-problems in any consistent fashion. All topics, whether those tested by E- or NE-problems, were covered in very similar fashions during lectures, recitations, and labs. Also, the fact that these performance differences developed and persisted through two different lecturers for 193/4 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 than I thought, and that their performance on E-problems 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.

Based on this analysis of alternative hypotheses, I feel that the evidence here best supports the conclusion that use of the evaluation tasks caused significant gains in student problem-solving performance on E-problems. It should be noted that the external validity of this claim is rather weak, though, since this study involved only two courses with very specific populations of students and learning environments. However, these results and their implications do seem to merit serious consideration by anyone interested in improving student problem-solving performance.

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. By comparing Figures 3 and 5, it appears that the students' use of special-case analysis related to their relative problem-solving performance, while there is no such apparent relation for unit analysis. Fortunately, there is a quantitative way to test this qualitative perception of the data.

To achieve this, we must first devise quantitative measures of students' overall use of each strategy on the exam evaluation tasks. In particular, we want the measures to best reflect students' apparent understanding of when, why, and how to use each evaluation strategy for a certain topic (i.e., the contextual strength of their evaluation meta-schemas).

The measures constructed are:

$$SCA_{relative} = FS_{193/4} \cdot QS_{193/4} - FS_{203/4} \cdot QS_{203/4}$$

$$UA_{relative} = FU_{193/4} \cdot QU_{193/4} - FU_{203/4} \cdot QU_{203/4}$$

where FS is the fraction of the class 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 were reported in Table 1.

I use FS and FU 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 $FS \cdot QS$ and $FU \cdot QU$, 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 quantify relative student performance on E- and NE-problems, I averaged the normalized difference in class performance for each problem category on each exam:

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

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

where ND represents the normalized difference between the two classes' performance (as plotted in Figure 5), and N denotes the number of problems of a certain type (either E- or NE-problems) on an exam.

Table 4 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 problem-solving performance, I conducted a correlation analysis of these data, as shown in Table 5. These results indicate that $E_{relative}$ is significantly positively correlated with $SCA_{relative}$ and uncorrelated with $UA_{relative}$, while $NE_{relative}$ is uncorrelated with both $SCA_{relative}$ and $UA_{relative}$. We also find that $SCA_{relative}$ and $UA_{relative}$ have a suggestively (though not significantly) strong negative correlation.

Exam	$SCA_{relative}$	$UA_{relative}$	$E_{relative}$	$NE_{relative}$
1	NA	NA	NA	-.102
2	.179	2.047	-.045	-.070
3	NA	NA	.024	-.037
4	.645	1.272	.066	-.098
5	.851	NA	.141	.010
6	.683	.893	.057	-.080

Table 4. Measures of relative overall class performance for exams 1 through 6. The definitions of each measure are described in the text.

	$SCA_{relative}$	$UA_{relative}$	$E_{relative}$	$NE_{relative}$
$SCA_{relative}$	1.000	-.966	.975	.447
	-	.166	.025*	.553
$UA_{relative}$		1.000	-.921	.528
		-	.255	.646
$E_{relative}$			1.000	.600
			-	.400
$NE_{relative}$				1.000
				-

Table 5. Pearson correlations between measures of relative class performance, and their p -values (2-tailed). * = significant at the .05-level.

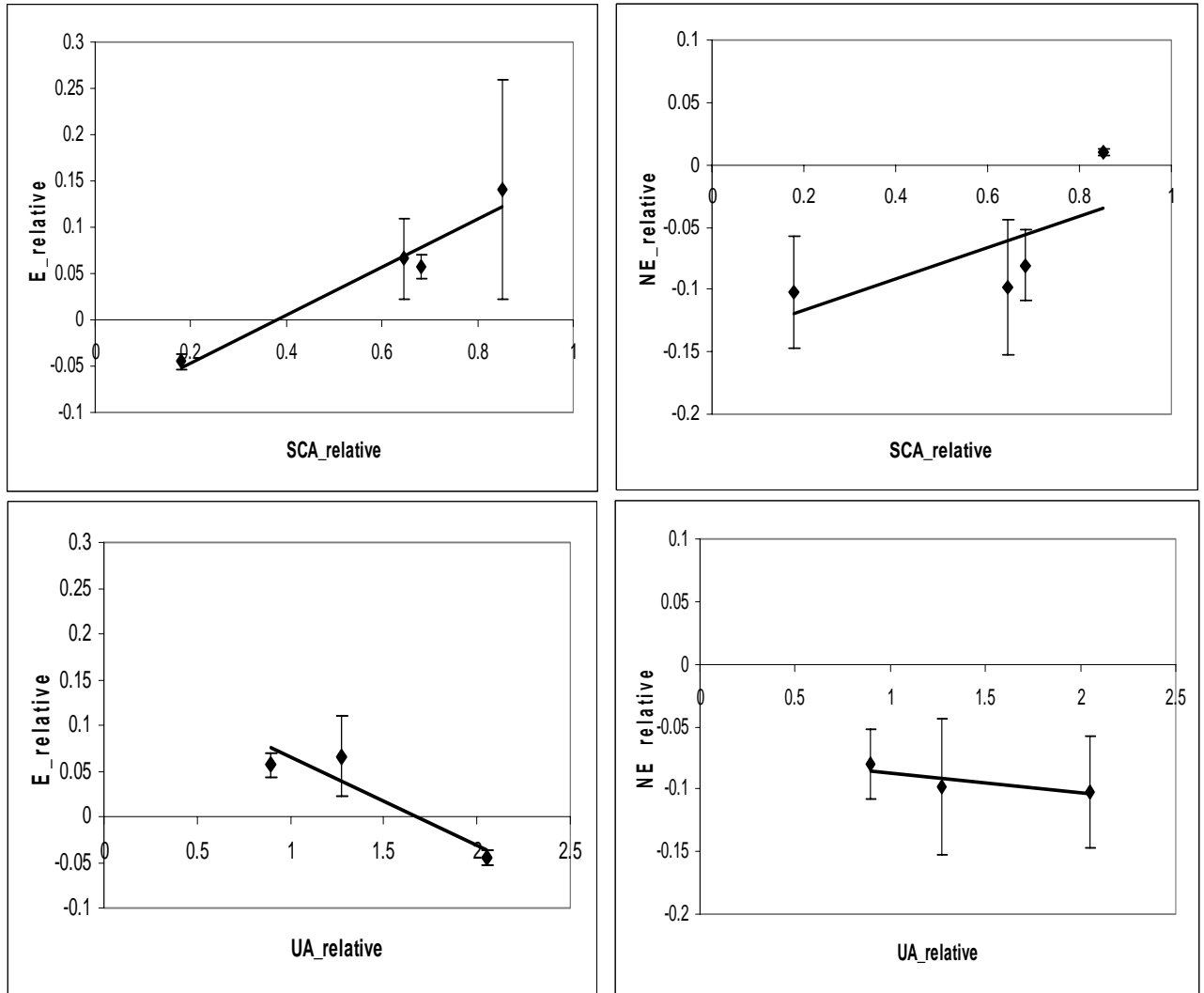


Figure 6. Plots of relative E- and NE-problem solving performance versus $SCA_{relative}$ and $UA_{relative}$, with trendlines and error bars included.

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) caused 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 1), and would consequently explain the strong negative correlation between $SCA_{relative}$ and $UA_{relative}$. It is somewhat disappointing, but not surprising, that students did not continue to use unit analysis as well as they had on exam 2. Above all, this illustrates the context-dependent nature of even a simple strategy such as unit analysis, and implies the importance of incorporating educational activities which help students understand how a general strategy is manifested in each context.

My 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. As Table 5 shows and Figure 6 illustrates, the relative performance of the 193/4 students on E-problems 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 (since such a factor should affect the problem-solving performance of both groups in a roughly equal manner).

The third hypothesis is that the use of unit analysis (in recitations, homework, and personal learning behavior) did not benefit students' problem-solving ability, or at least any such benefit was much weaker than the benefits due to special-case analysis. This is certainly consistent with Table 5 and Figure 6. It would appear that the 193/4 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 2). 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.

The fact that $NE_{relative}$ is uncorrelated with both $UA_{relative}$ and $SCA_{relative}$ supports our assumption that there is no transfer in the benefits of either strategy to topics not covered by the evaluation tasks from recitation and homework assignments. It may be that such transfer is possible if student motivation and self-efficacy are strong enough, something which studies involving physics majors or even graduate students may be able to address.

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 problem-solving benefits. In contrast, the simpler strategy of unit analysis is learnt very quickly, but apparently provided little or no problem-solving benefits. Here we shall discuss some possible reasons for these results.

In his doctoral dissertation, Sherin²⁸ 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 6. There are three classes of

devices; narrative, static, and special case. Devices in the narrative class function by imposing an imaginary change in the physical system modeled by the equation and then examining the resultant behavior of the equation. Static devices instead focus on studying the equation at some distinguished instant during a physical process. Special case devices restrict the values of quantities in the equation in order to study its behavior.

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

Table 6. Interpretive devices listed by class. Reproduced from Sherin [28] (Chapter 4, Figure 2).

Each of these devices plays a role in what we call special-case analysis (note that I use the term “special-case” in a much broader sense than Sherin when I speak of “special-case analysis”). One may conduct many different special-case analyses of an equation, using any one of these interpretive devices as the specific means for utilizing the analysis strategy. 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, we therefore require students to utilize 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 we make meaning using resources such as systems of words, images, actions, and symbols. Semiotics research has identified two aspects of meaning, called typological meaning (i.e., meaning by kind) and topological meaning (i.e., meaning by degree)²⁹. 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 activities³⁰.

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. The motion is now constant velocity linear motion, and we find that a is zero which is consistent with this fact. Therefore, this aspect of topological meaning construction also serves to develop typological meaning by highlighting the necessary role that $a=v^2/r$ plays in circular motion. When given a problem about circular motion in the future, the student may now be more likely to remember to use this equation for acceleration.

I argue that interpretive devices are a primary means by which topological meaning is constructed for an equation, and have the ancillary effect of developing typological meaning as well. So 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 problem-solving performance. Unit analysis, on the other hand, makes no 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 is far weaker than special-case analysis at developing meaning for equations. When students use unit analysis, it is usually employed as an algorithm, not as a means for conceptually examining the equation. 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. However, it seems that unit analysis does not carry much worth as a means for improving student problem-solving performance,

and may not provide any significant benefit to student understanding of the subject matter.

Conclusions

Evaluation is a well-recognized part of learning, yet its importance is often not reflected in our introductory physics courses. I therefore have developed tasks and rubrics designed to help students become evaluators by engaging them in formative assessment activities. By incorporating evaluation tasks into recitation and homework assignments, I believe that students gain a greater degree of empowerment as they are better able to learn and do physics on their own. In particular, students can check and correct their own work when solving problems, at least to some degree. Additionally, special-case analysis tasks engage students in the use of interpretive devices to construct typological and topological meaning for equations. This is, however, a new avenue of research in physics education and there is a great deal of work to be done. Future studies designed to replicate and expand upon these results are clearly 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.

Acknowledgements

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, I would like to thank the National Science Foundation (DUE-0241078) for funding which partially supported my work.

¹ E. Etkina, A. Van Heuvelen, D. Brookes, S. Murthy, D. Rosengrant, A. Warren, Phys.Rev.ST-PER (submitted).

² B.S. Bloom, *A Taxonomy of Educational Objectives, Handbook I: The Cognitive Domain*. David McKay Co Inc, New York (1956).

³ *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*, edited by L.W. Anderson & D. R. Kraftwohl, Longman, New York (2001).

⁴ A.E. Lawson, *Science & Education* **9**, 577-598 (2000).

⁵ A.E. Lawson, *The American Biology Teacher* **62** (7), 482-494 (2000).

⁶ A. Buffler, S. Allie, F. Lubben, B. Campbell, *International Journal of Science Education* **23** (11), 1137 (2001).

⁷ R.F. Lippmann, Ph.D. Dissertation, University of Maryland (2003).

⁸ S.P. Marshall, *Schemas in Problem-Solving*, Cambridge University Press, NY (1995).

⁹ M. Sabella, E.F. Redish, *Knowledge organization and activation in physics problem-solving*, University of Maryland pre-print (2004).

¹⁰ R.R. Hake, *Am. J. Phys.* **66**, 1 (1998).

¹¹ K. Cummings, J. Marx, R. Thornton, D. Kuhl, *Am. J. Phys.* **67**, S38 (1999).

-
- ¹² N.D. Finkelstein, S.J. Pollock, *PRST-PER*, **1**, 1 (2005).
- ¹³ A.P. Fagen, C.H. Crouch, E. Mazur, *Phys. Teach.* **40**, 206 (2002).
- ¹⁴ R. Warnakulasooriya, D.J. Palazzo, D.E. Pritchard, *Evidence of Problem-Solving Transfer in Web-Based Socratic Tutor*, in *Physics Education Research Conference Proceedings*, edited by P. Heron, L. McCullough, J. Marx, AIP Press, Melville (2006).
- ¹⁵ K. VanLehn, C. Lynch, K. Schulze, J. Shapiro, R. Shelby, L. Taylor, D. Treacy, A. Weinstein, M. Wintersgill, *International Journal of Artificial Intelligence in Education*, **15** (3) (2005).
- ¹⁶ B.J. Zimmerman, M. Martinez-Pons, *American Educational Research Journal* **23**, 614-628 (1986).
- ¹⁷ N. Perry, *Journal of Educational Psychology* **90**, 715-729 (1998).
- ¹⁸ D. Hammer, *Am. J. Phys.* **64**, 1316-1325 (1996).
- ¹⁹ American Association for the Advancement of Science, Project 2061, *Benchmarks for science literacy*. Oxford University Press, New York (1993).
- ²⁰ 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/du/documents/review/96139/four.htm> (1996).
- ²¹ National Research Council, *National science education standards*. National Academy Press, Washington, D.C. (1996).

-
- ²² R.H. Ennis, *A taxonomy of critical thinking dispositions and abilities*. In J.B. Baron and R.J. Sternberg (Eds.), *Teaching Thinking Skills: Theory and Practice*. New York: Freeman. P. 9-26 (1987).
- ²³ P.M. King, K.S. Kitchener, *Developing Reflective Judgment: Understanding and Promoting Intellectual Growth and Critical Thinking in Adolescents and Adults*. San Francisco: Jossey-Bass. (1994).
- ²⁴ P. Black and D. Wiliam, *Inside the black box: Raising standards through classroom assessment*, London: King's College, (1998).
- ²⁵ R. Sadler, *Instructional Science* **18**, 119-144, (1989).
- ²⁶ A.R. Warren, *The Role of Evaluative Abilities in Physics Learning*, 2004 Physics Education Research Conference Proceedings, Sacramento, CA, edited by J. Marx, S. Franklin & P. Heron, American Institute of Physics, (2005).
- ²⁷ A.R. Warren, Ph.D. Dissertation, Rutgers University (2006).
- ²⁸ B. Sherin, Ph.D. Dissertation. University of California, Berkeley, 117-134 (1996).
- ²⁹ J.L. Lemke, "Multiplying Meaning: Visual and Verbal Semiotics in Scientific Text," in *Reading Science*, edited by J. R. Martin & R. Veal (Routledge, London, 1998).
- ³⁰ E. Etkina, A.R. Warren, M.J. Gentile, M. J., *Phys. Teach.* **44**, 1 (2006).