

**What happens between pre- and post-tests: multiple measurements of student understanding during an introductory physics course.**

Andrew F. Heckler<sup>a)</sup> and Eleanor C. Sayre<sup>b)</sup>

<sup>a)</sup> Department of Physics, Ohio State University, Columbus, Ohio 43210

<sup>b)</sup> Department of Physics, Wabash College, Crawfordsville, Indiana 47933

Abstract

Pre- and post-testing cannot measure peaks or troughs in student performance, nor does it indicate when performance changes during a course. To better characterize the evolution of student understanding, we posed simple conceptual questions several times per week to separate, randomly selected groups of introductory physics students throughout the course. This design avoids issues of retesting and allows for the tracking of student understanding of a given topic during the course with a time resolution on the order of days. Reporting on findings from data collected from over 1600 students over five quarters, we found significant and interesting changes in performance on timescales of days and weeks. We found that “response curves” of performance vs. time can be divided into distinct shape categories. The first is flat, representing no change, due to either ceiling or floor performance or ineffective

instruction. The second shape is a rapid step-up in performance during instruction, with no subsequent change. The third is a rise due to instruction and decay, possibly due to memory decay. The fourth is a rapid decrease in performance, likely due to interference from learning a related topic. Finally, we examined changes on a one-day time scale and found that changes in performance did not coincide with relevant traditional lectures or labs, but do coincide with due dates of relevant on-line homework that provides immediate feedback. This method is well-suited to measuring the effect of particular instructional events as well as providing insight into the nature of student understanding and learning.

## **I. INTRODUCTION**

How much does student understanding of a particular physics topic change over the course of instruction? To answer this question, instructors and education researchers have commonly administered pre- and post-tests to students at the beginning and end of a course or unit. However, measuring the changes in student understanding *between* the pre and post tests can also be of significant interest to both instructors and education researchers. For example, measuring understanding throughout a course can help to address questions such as: does learning of a particular topic happen during lecture, homework, or some other part of the course? Are there rapid rises and decays in understanding? Does learning one topic interfere

or help with understanding of another related topic? These kinds of question can be better answered by measuring student understanding repeatedly on timescales of days and weeks throughout the course.

In this paper, we report on the evolution of student responses on a number of simple conceptual questions posed several times per week throughout an introductory physics course to separate, randomly chosen groups of students enrolled in the course. We present this data in the form of “response curves”, namely graphs of performance on given questions as a function of time. In a recent paper<sup>1</sup>, we identified some interesting features such as peak-and-decay and temporary interference in a few select response curves from a portion (representing one quarter) of the data presented here. Here, we report on a larger collection of data from 10 classes (~1700 students, 5 quarters) in order to more systematically present and highlight general results from the study.

We begin by describing the background and rationale for the study and the details of the experimental design. In sections IV-VII, we present examples of a number of response curves, categorize these curves into four basic shape categories, and briefly discuss the possible mechanisms responsible for the basic shapes. In Section VIII, we present examples of response curves in higher resolution (daily averages) and discuss the timing of features in the curves with instructional interventions. Finally we conclude with a general discussion of the results.

## **II. BACKGROUND AND RATIONALE**

The goal of this study is to measure and characterize general patterns of changes in student understanding of a select set of physics topics throughout a course. The term “understanding” is quite ambiguous; in this paper we will operationally define a measurement of student understanding as the performance on simple conceptual multiple choice problems. Clearly this definition has its limitations, as it does not include the richness of dimensions that can describe the complex nature of understanding and performance. Some formal methods such as the microgenetic method<sup>2,3</sup> employ frequent quantitative and qualitative measures of complex learning processes, such as problem solving. However these methods are limited in practice to a small number of students. In contrast, multiple choice tests may be administered to large numbers of students, allowing for the construction of statistically reliable response curves for a large population of students.

The validity of the simple multiple-choice questions used in this study were supported in a number of ways. First, after the questions were administered we conducted informal, short post-interviews with the participants to verify that the questions were understood as intended and that their answers were interpreted appropriately. Second, to the extent that validity is partially supported from correlations with other measures of relevant knowledge (criterion and construct validity) most of the questions used in this paper were significantly correlated with final course grade ( $p < 0.05$ ). Further, most of the response curves here show features

coinciding with relevant instruction, lending further support to construct validity. Finally, a number of these questions were derived from questions in more formally validated assessments such as CSEM<sup>4</sup> and DIRECT<sup>5</sup>, or are questions commonly found in lectures or the “conceptual question” section in the end of chapter questions.

Remarkably, we have found no other studies continually tracking understanding of populations of students on timescales of days or weeks during instruction. Most of the previous studies on the evolution of student knowledge can be described as either studies of the long-term retention of knowledge after the end of instruction and/or studies involving pre- and post-tests a few days or weeks before and after instruction. Outside the realm of physics education, there have been a number of studies on long-term retention, typically on the order of a few weeks to a few years after instruction, extending to up to 50 years, as summarized in extensive reviews<sup>6,7</sup>.

In the physics content area, there are a few small scale empirical studies of student retention<sup>8,9,10</sup> or short term changes in student understanding (e.g., Ref. 11, 12), and some discussion about models and methods of measuring changes in understanding<sup>13,14,15</sup>. The vast majority of studies of changes in student understanding in physics education involve pre- and post-testing, usually to compare the effectiveness of instructional methods (e.g., Ref. 16). Ding et al.<sup>17</sup> have found that pre- and or post-tests can change significantly by simply varying the day on

which it is administered. They identified changes on particular questions relevant to the lectures just before the test as the cause, thus highlighting the importance of changes in understanding at smaller timescales.

There are a number of fundamental reasons to expect interesting changes in student performance on time scales of the order of days and weeks. First, since forgetting of knowledge occurs on all times scales, with the largest loss occurring just after learning<sup>18,19,20</sup>, it is possible that much knowledge learned in a physics class may be forgotten within hours or days after learning. Second, while learning complex skills such as problem solving may take weeks or even years, learning simple concepts such as the idea that an electric field exists around a charged particle, could happen quite rapidly. Finally, the learning of one topic may interfere with understanding or learning another topic, as is well documented in memory studies<sup>21,22</sup>. Since learning of a topic may happen quickly, interference may happen quickly as well.

### **III. EXPERIMENTAL DESIGN**

One major challenge in measuring the evolution of student understanding is the potential for any unwanted effects of retesting students. For example, in simple memory tasks, students who were tested between training and a final test, even with no feedback, tended to score higher compared to those who spent time studying the material instead of testing<sup>23,24</sup>. While more complicated physics conceptual tests may

have minimal retesting effects when the testing is separated by many weeks<sup>25</sup>, it is not unreasonable to assume that taking the same test twice within a period of a 1-3 days and/or many times during a course may introduce retesting effects.

To avoid any potential effect of retesting, the design is a between-student cross sectional study rather than a within-student longitudinal study. During each course, we systematically cycled through all of the students once by quasi-randomly selecting a group of students for testing each week of the course. Each randomly selected group was considered as statistically equivalent, allowing for comparison of performance between groups throughout the quarter.

In order to verify that the groups were reasonably equivalent in ability, we performed for each question an ANOVA test comparing average final course grades of the groups for each week and found no significant differences between groups (all  $p$ -values  $> 0.10$ ). This equivalency allowed for comparison of weekly averages using a simple chi-squared analysis. However, to further verify that the results were not biased by groups with higher grades, we have also performed statistical analysis for each question (when possible) using a general linear model, including *week* as a factor and *final grade* as a covariate. The results increased or decreased  $p$ -values by only small amounts, not enough to alter our conclusions about the statistical significance of the observed features. Furthermore, the response curves reported here have obvious features (large effect sizes), and none of the curves of average final course grade vs. time or response curves of other test questions on unrelated

topics appear to follow the same pattern. In short, the response curves reported in this paper do not appear to be explainable by differences in average grade between the groups or average performance on other unrelated questions. Thus our assumption that the groups were equivalent prior to instruction seems to be reasonable.

### **A. Participants and method**

The participants in this study were university students enrolled in a first quarter (Mechanics) or second quarter (Electricity and Magnetism) introductory calculus-based physics course primarily designed for engineering students. Each lecture section had a typical enrollment of 170, and we studied two lectures sections each quarter for 5 quarters during the academic years 2007-8 and 2008-09. In total, 1694 students participated over 5 quarters, with roughly equal numbers of student each quarter. For each quarter, the instructor was a regular physics faculty member who taught both sections. This represents 4 different faculty; three taught the course a number of times previously, and one was a first-time teacher of the course. The instructional method was a traditional lecture format, with no interactive methodologies (such as voting machines) used beyond the occasional answering of a question from a student. In addition to the 3 lectures per week, there was one recitation and a traditional lab with “cookbook-like” confirmatory lab activities.

Homework was assigned via the online platform WebAssign<sup>26</sup> once per week and consisted of ~10 typical end of chapter problems<sup>27</sup>. The answer format was usually typed in numerical answers or multiple-choice with no work shown. Immediate feedback was given for answers and students could have up to 10 attempts without penalty to input the correct response.

In addition to the standard homework, students were also given a “flexible homework” assignment as part of their regular course credit (for participation). The flexible homework assignment consisted of participating in a one-hour session in our physics education research lab where students would complete some combination of training, testing, and interviewing. Data reported in this study are from these sessions. Each week during the course we would randomly select 1-2 lab sections (out of a total of about 14), and ask students to sign up for flexible homework. Typically, about 95% of students participated in the flexible homework and all of these students participated in the study.

During the flexible homework session, students were told to answer the questions as best they could, even if they have not yet seen the material. They sat at individual stations in a quiet, proctored room to answer several series of physics questions either with pencil and paper or on the computer. Students completed the material at their own pace. Afterwards we would informally ask students whether they had any questions and/or to explain their answers. We observed during these

sessions that students made an effort to answer the questions to the best of their ability.

In addition to the data collected from the flexible homework session, we also collected materials such as the syllabus, lab book, and homework assignments in order to determine more precisely when relevant events in the course, such as a lecture or lab on a particular topic occurred. Lectures were observed, and field notes on lecture content were recorded, including the approximate level of attendance. Finally, we also collected the grades for each student in the course.

## **B. Analysis of data**

As stated earlier, the data collection method in this study is novel, thus one main goal of this paper is to report on any general patterns and trends in this new kind of data. We used a straightforward three-phase strategy to look for patterns. The first phase was to bin the data for each question by day, week, and two-week periods and perform a chi-squared test of independence for each binning. This was a critical first test to determine whether there was *any* significant variation in student performance for each test item. A total of 126 test items were used over the course of the study, and we found that there was significant variation in performance on only 37 (about 30%) of the items. We also found that the binning did not qualitatively change this result. For all but two of the questions reported here the patterns were replicated in more than one quarter. The two that were not replicated (Figures 3 and

5) were administered in only one quarter, and the signals were very large and significant. As stated earlier, this data comes from 4 separate traditional lecturers. Since the goal of this paper is not to compare instructors but rather to characterize general kinds of patterns in student performance, we did not distinguish between instructors. Nonetheless it is interesting to note that the performance on specific questions between classes with different instructors was largely similar. Clearly this method could be used to design more focused studies comparing instructional methods.

The second phase of data analysis consisted of an examination of the response curves of the questions with significant variation (as determined by Phase 1) to determine whether the significant variations occurred during relevant instructional events and to determine the general shape of the variation. It is significant to note that all of the variations found coincided with relevant instructional events; we found no curves with sudden significant increases or decreases that were not coincident with relevant instruction events. It is in this phase that we noticed four general categories of response curve shapes (flat, step-up, rise and decay, and step-down, as described in subsequent sections), and these shapes appear to coincide well with the general causal mechanisms of learning, decay, and interference mentioned above.

The third phase consisted of straightforward statistical tests to determine whether the response curve features were consistent with simple learning, rise and

decay, or interference. Simple learning (step-up in performance) was tested using a chi-squared test for independence, comparing performance before and after relevant instruction. Testing for significant rise and decay also used the chi-squared test to compare performance before and just after instruction *and* to compare performance between just-after and long-after instruction. Testing for interference (step-down in performance) was similar to simple learning. Because the features of the response curves were quite distinguishable, many of them could also be verified as statistically reliable by simply inspecting the response curves including error bars.

Finally, we would like to note that while the analysis used in this paper is valid for the purposes of a simple characterization of the general response curve patterns, a more precise quantitative analysis requires a more detailed model for the shape of the curves, such as a simple model for learning and memory decay based on cognitive psychology models briefly described in our previous work<sup>31</sup>. Such a quantitative analysis, for example using Maximum Likelihood Estimation methods<sup>32</sup>, allows for more precise quantitative parameterization, comparisons and hypotheses testing. Such an analysis is beyond the scope and goals of this paper and is certainly a fertile area for more in-depth investigation.

### **C. Comment on non-longitudinal data**

One must be careful in making inferences about the evolution of individuals from non-longitudinal data. In particular, one must keep in mind that the shape of the

response curves here represent the evolution of the *average* of the population, and this does not necessarily imply that individuals follow this same path. For example, a “rise and decay” population average response curve could be comprised of two sub-populations, one that steps-up in performance while the other is flat, then decreases. Without a longitudinal study of individuals, there is no way of being certain how particular individuals or subpopulation of individuals evolve<sup>28,29</sup>.

Nonetheless, from a larger perspective, cross sectional data, such as the evolution of the class average, can certainly produce useful information to instructors and education researchers alike, and some conclusions can still be made about the evolution of individuals. For example, one can rule out the hypothesis that a significant population of students individually follow a given shape if the response curve does not follow this shape. One can also make useful and reasonable assumptions about factors that may separate out sub-populations of students into different response curves, such as math ability or final grade, and increase one’s confidence that a significant number of students in a given sub-population are following or not following a given path<sup>30</sup>.

#### **IV. FLAT CURVES: PERFORMANCE AT CEILING OR FLOOR?**

As stated earlier, many (about 70%) of the conceptual questions posed resulted in no significant variation in student performance over the quarter. In other words, the data was an unchanging (smooth and flat) response curve. This is

consistent with several seminal physics education research studies finding a lack of change from pre- to post-test for many simple conceptual questions (e.g. Ref. 33, 16). However the lack of difference between pre- and post-test scores does not preclude the possibility that temporary peaks may occur during the quarter. In any case, it can be somewhat striking to see smooth, flat curves, i.e. no change at all on a given question over the course of the quarter, even during instruction relevant to the question.

Figure 1 and Fig 2. present examples of flat response curves and corresponding questions. For these and most of the other graphs in the paper, the time window of relevant instruction, which includes lecture, homework and lab, is also indicated. The questions in Fig. 1 are part of an instrument in development, and the questions have been shown to be reasonably valid and reliable<sup>34</sup>. The first question in Fig.1 tests for the known misconception that the net force on an object must be in the same direction as its motion. The score on this question remains unchanged throughout the quarter ( $\chi^2(7) = 7.0, p = 0.43$ ). The second question in Fig. 1 is a relatively easy question about velocity and acceleration, and correct responses remain unchanged throughout the quarter ( $\chi^2(7) = 2.6, p = 0.90$ ). The question in Fig. 2 is a simplified version of a question from DIRECT, an instrument for assessing understanding of direct current circuits<sup>35</sup>. The scores on this question do not change over the quarter ( $\chi^2(9) = 6.0, p = 0.74$ ).

The upper curve in Fig. 1 with an average score of 83% is likely at ceiling, and the lower curve with an average of 14% is likely at floor. It may also be the case that the question in Fig. 2, with an average score of 47%, is also scoring near floor. Considering three possible answer choices for this question, and only 15% of students chose “equal”, this leaves most students choosing between one of two answers, “brighter” or “dimmer”, and the proportion choosing each remained constant throughout the quarter. Thus, it is possible that most students were randomly guessing one of these two choices, resulting in about half getting the problem correct throughout the quarter. It is also possible that half of the students always knew the correct answer throughout the quarter, though from our informal debriefings with students, this seems highly unlikely. This finding that flat curves were either at floor or ceiling was consistent throughout our study.

## **V. STEP-UP**

The next response curve shape presented is the “step-up” in performance, characterized by an initial period of no change in performance followed by a rapid increase to a new level of performance which is maintained for the remainder of the course. It is worth noting that all step-up curves observed coincided with a relevant instructional event, such as a relevant homework, and there were no step-ups observed that did not coincide with a relevant instructional event. Here we present two step-up curves. Figure 3 presents the response curve for a basic question that one

might find in a typical lecture on electric field or in the conceptual questions section in the back of a chapter in a textbook<sup>36</sup>. The weekly average proportion of correct responses does change over the course  $\chi^2(9) = 38.6, p < 0.0001$ , with a clear step-up shape showing a significant difference between the average scores before (9%) and after (54%) instruction,  $\chi^2(1) = 33.9, p < 0.0001$ , Cohen's effect size:  $d = 1.1$ . The most popular incorrect response is "a", due to the common misconception that the contribution from the closer plate is greater.

Figure 4 probes well-known difficulties students have with understanding circuits<sup>37,38</sup>. This curve displays a significant difference in performance during the course  $\chi^2(9) = 20.9, p = 0.01$ , with a significant step up during instruction showing a significant difference between the average scores before (19%) and after (41%) instruction,  $\chi^2(1) = 16.1, p < 0.0001, d = 0.5$ .

## **VI. RISE-DECAY ON SCALE OF WEEKS**

As mentioned earlier, the decay of memory is well established at least for some kinds of knowledge. To the extent that a particular topic is addressed only for a limited time during the course, it might be expected that student performance would rise during instruction and decay afterward. We present two possible examples of this. Figure 5 presents a rise and decay curve resulting from a simple question about electric fields. The curve significantly changes during the course  $\chi^2(9) = 28.1, p = 0.001$ , with a clear peak during instruction, and subsequent decay. As to be expected,

the most popular incorrect answer is ‘c’(the field is greater near the infinite plates), which most students chose overwhelmingly at the beginning and end of the course, but not as much during instruction. This may be an example of a ‘misconception’ returning after instruction.

Figure 6 presents a rise and decay curve, though the curve is more complicated. First, there is another rise in the last week. This is likely due to a review homework assignment at the end of the quarter which included similar questions. This will be discussed more in Section VIII. Second, while the performance does change significantly during the course  $\chi^2(8) = 16.2, p = 0.04$ , there is reason to believe that the “decay” in performance is in fact due to explicit interference with a related topic. We address this issue of interference, including this particular example, in the next section.

## **VII. INTERFERENCE**

When two topics are perceived as related, the learning of one may interfere with the learning and memory of the other. For example, an abundance of evidence suggests that a major cause of forgetting a particular piece of knowledge is simply the learning of new knowledge, especially if the new knowledge is in some way similar<sup>39,40</sup>.

We present two examples of interference. The first example involves vector and scalar quantities in electromagnetism. We have found that students often fail to

recognize the importance of keeping in mind the vector nature of the electric field  $E$  and the scalar nature of the electric potential  $V$ . This confusion may occur because the concepts of electric field and electric potential, both unfamiliar, abstract quantities used in electrostatics, may be perceived as to two highly similar quantities by students, and as a result  $E$  and  $V$  they may interfere with each other in the course of learning and problem solving. Thus when students are asked questions about  $E$  or  $V$  for particular charge distributions, they often seem to treat either of them as a vector or a scalar, depending on the question<sup>41</sup>.

In a recent paper<sup>42</sup>, we provided evidence that learning about the electric potential, interferes with student understanding of the vector nature of an electric field. The first example of interference in this paper in Fig. 7 replicates this finding and extends it to demonstrate that not only does learning about electric potential interfere with the understanding of electric field, but also learning about electric field (and magnetic field) can interfere with understanding of electric potential. In particular, as shown in Fig. 7 the students correct “vector-like” responses to  $E$ -field questions increased during instruction about  $E$ , then quickly turned to the scalar-like responses during the instruction on  $V$ , then returned to vector-like responses after the instruction of  $V$  ended and the instruction on the vector magnetic field began. The responses marginally change over the course,  $\chi^2(16) = 25.2$ ,  $p = 0.065$ , and the average vector response is higher during electric and magnetic field instruction (61%) than during electric potential instruction (46%),  $\chi^2(1) = 5.8$ ,  $p = 0.02$ ,  $d = 0.3$ .

Conversely, as shown in Fig. 6, during electric field instruction, students begin to use (incorrect) vector-like responses, then turn to (correct) scalar-like responses during electric potential instruction, then return to vector-like responses during magnetic field (vector) instruction. The responses significantly change over the course  $\chi^2(16) = 36.2, p < 0.01$ , and the average scalar response is higher during electric potential instruction (59%) than during electric and magnetic field instruction (43%),  $\chi^2(1) = 7.5, p < 0.01, d = 0.3$ .

In short, Figures 6 and 7 indicate that a significant number of students answer both E and V questions as vector-like during vector field instruction (E and B) and answer as scalar-like during scalar field instruction (V). While the effect size ( $d = 0.3$ ) is somewhat small, it appears to be reliable, as we have seen the effect in two separate quarters.

The second example of interference involves the electric force and the magnetic force. Since the representation of vector E fields and B fields are often similar, and both are invisible fields which exert a force on a charged particle, one might expect that students could confuse the electric force on a charged particle with the magnetic force. Figure 8 shows the evolution of student responses to a simple question about the direction of the magnetic force on a charged particle. Before magnetic force instruction there were a significant number of students (57%) answering (incorrectly) that the force is in the direction of the magnetic field, especially just after E-field instruction<sup>43</sup>. Following this, there was a clear rise in the

correct responses during/after magnetic force instruction (63%) compared to before instruction (14%),  $\chi^2(1) = 66.1, p < 0.0001, d = 1.1$ . Consistent with previous findings<sup>44</sup>, this indicates that the learning of electric force may interfere with students' understanding of magnetic force in that they initially assumed that a magnetic field exerts a force on a (positively) charged particle in the direction of the field, just as an electric field would, since they have not been taught otherwise. Once taught, many students learn magnetic force quickly.

The more dramatic signal of interference<sup>45</sup> comes from the response curve of an equivalent question about the electric force, as shown in Fig. 9. Early in the quarter, students answered that the electric force is in the direction of the field as they were taught, and 65% answered correctly. However, once magnetic force was taught, there was a sudden decrease to 44% of answers in the direction of the field, and a large rise from 11% to 43% in answers *perpendicular* to both the velocity  $v$  and  $E$ , similar to the magnetic force which is perpendicular to  $v$  and  $B$  ( $\chi^2(1) = 34.8, p < 0.0001, d = 0.8$ ). This is a strong indication that learning about the magnetic field force interfered with students' understanding of electric field force. It is interesting to note from comparing Fig. 9 and Fig. 8 that movement away from the correct electric force answer only occurs about 1-2 weeks after there was significant learning of the B-field force.

## VIII. RAPID CHANGES

The previous sections presented changes in student performance on a timescale of weeks. However, since data was typically collected several times per week, we can observe changes on the timescale of days. This increased time resolution can help us to answer a number of new questions, including whether performance changes on the scale of days and if so, whether sudden increases in performance coincide with the relevant lecture, homework, lab, or recitation. Here we present two curves that show clear features.

The first example, shown in Fig. 10 is a question about circuits that is closely related to a question in DIRECT. There are two notable features of this response curve. First, there is a rapid rise in performance, and this rise does not coincide with the relevant lecture or labs, but rather coincides with a relevant homework set. In particular, the rapid rise begins on quarter day 38 and the relevant lecture occurred on quarter day 30 (attendance ~60%) in which an explicit example was presented with the same combination of resistors and also included a closely related demonstration. Further, there were two weeks of labs on circuits including multiple loop circuits, on quarter days 28-30 and 33-35. The relevant homework assignment was due on quarter day 39 and included three problems on multiple loop circuits, including one very similar to the question in Fig. 10. Since the homework was online, we were able to obtain electronic records of students' homework activity, and

we found that over 60% of the students completed the relevant problems within two days of the deadline. In particular, comparing the performance before the relevant homework, from days 1-37, to the performance from days 38-43 (during and just after when homework was due), there was a significant difference in performance ( $\chi^2(1) = 29.6, p < 0.001, d = 0.86$ ). Thus it is likely that the homework is the cause of the large and rapid rise in performance and not traditional lecture or lab.

The second notable feature of the response curve in Fig. 10 is the decrease of performance in the time of several days to one week. In particular, the average score from days 38-43, (during and just after the homework was due) was 66% which is significantly greater than the score from days 44-49, several days to one week after homework ( $\chi^2(1) = 3.8, p = 0.05, d = 0.37$ ). This is consistent with our previous finding suggesting that student performance can decrease significantly even within one day<sup>46</sup>. A more precise parameterization of the decay time would require of more careful modeling, as mentioned in Section III. B.

The second example, in Fig. 11, presents a response curve for the average of a collection of 4 similar questions about the electric potential resulting from two point charges of the same sign or opposite signs. The results are similar to the first example in that there is a sudden rise in performance that coincides with the homework and not with lecture or lab. In particular, there is a distinct rise in performance on quarter day 20, while the relevant lectures (attendance ~70%) occurred on quarter days 16 and 18, the relevant labs occurred on days 14-18, and

the relevant homework was due on quarter day 21, and we found that more than 65% of the students completed the relevant homework problems less than two day before the homework was due. In particular, comparing the performance before the relevant homework, from days 1-19, to the performance from days 20-25 (during and just after when homework was due), there was a significant difference in performance ( $\chi^2(1) = 30.5$ ,  $p < 0.001$ ,  $d = 0.8$ ). Two other features are worth noting. The first is the dip in performance around quarter day 42. This is likely due to the interference effect discussed in a previous section, as the vector-like responses peaked at this time. Finally, there is an increase in performance on quarter day 48. This day coincides with a review homework assignment in which a question about the electric potential from two charges is asked. Thus, it appears that this relative increase in performance is due to the review homework.

In sum, these two examples provide striking evidence that the increase in performance was not due to what students learned in the traditional lectures, rather what they learned by doing homework with immediate feedback.

## **IX. CONCLUDING REMARKS**

We have found significant changes in student scores on simple conceptual questions occurring on both the day and week timescales over the term of a course. We represented the results as response curves of average score verses time and found four basic shapes with associated simple causes for each shape.

The first shape is flat, indicating no change even when change might be expected from instruction. The flat curves in this paper and other flat curves found in the study appear to be at ceiling or floor. This could be simply explained by the possibility the population was significantly above ceiling or below floor. Another simple explanation for flat curves is that the instruction itself is simply ineffective, and finding flat curves only at ceiling or floor is purely coincidence. Whether the flat curves are due to the nature of the question or instruction or both remains an open question.

The second shape of response curves found is the step-up shape, with a sudden increase often on the scale of days, coinciding with an instructional event and with a plateau lasting the remainder of the course. Because forgetting is such a ubiquitous phenomenon it may be somewhat unexpected for there to be an (apparently) unchanging plateau after instruction. This may be due to a long decay time, or to the possibility that the topic is constantly practiced at some level, and the plateau presumably represents a ceiling in performance.

The third shape is a relatively rapid increase followed by significant decrease in score. This peak can happen on the time scale of days or weeks. This highlights the fact that student performance can be quite dynamic, changing both up and down fairly rapidly over a range of timescales. The increases shown in this study were coincident with instruction, while the decreases may be due to forgetting, explicit interference from a related topic, or some other factor. The reason why some scores

decay in days, some in weeks and some not at all is subject to further study. One possibility for why a particular score rises and decays is that the students initially have a strongly held misconception which is only temporarily changed during instruction.

The fourth shape is somewhat related to the third: it is a rapid decrease in score that coincides with the learning of a related topic. For the examples provided in this paper, we argue that this shape is likely due to interference.

#### **A. Implications for instruction**

While this study only examines the evolution of scores on simple conceptual questions and does not address the evolution of, for example, problem solving skills, the results of this study have several implications for instruction beyond what has been already been learned from pre- and post-testing.

First, the potential for rapid changes in scores, especially in the form of peaks in performance indicates that pre- and post-testing does not always characterize the evolution of student learning adequately and may give misleading information to the instructor.

Second, given the knowledge that student understanding may rapidly peak for a particular topic, even if only momentarily, instructors may be able to adjust instruction to extend the peaks. For example, there are a number of studies indicating

that repeated and appropriately spaced practice can dramatically increase retention<sup>47,48,49</sup>.

Third, this method can provide evidence of the potentially dramatic effects of interference. Knowledge of how one topic, such as electric and magnetic fields can interfere with understanding of another similar topic, such as the electric potential field, can help instructors adjust instruction to address this issue. It is important to point out that evidence of interference is a specific example of the power of having high resolution information of the evolution of student models. Accurately measuring when the interference occurs aids in the identification of the specific interfering topic. Examining the dynamics of student models of understanding is certainly a rich topic for further study.

Finally, it can be very useful to accurately know when student understanding improves during the course, and what instructional events caused this to happen. In this paper, the increases in performance only happened directly after homework (with feedback), and not directly after traditional lecture or lab. This is yet another reason to reconsider the value of traditional lecture and lab. This design is well-suited to test the effect of a particular instructional method or curricular change used at a particular time in the course.

## **ACKNOWLEDGMENTS**

This research is partially supported by a grant from the Institute of Education Sciences, U.S. Department of Education (#R305H050125). We would like to thank Thomas Scaife and Rebecca Rosenblatt for help in designing questions and collecting data.

## **REFERENCES**

<sup>1</sup>E. C. Sayre and A. F. Heckler, "Peaks and decays of student knowledge in and introductory E & M course," *Phys. Rev. ST: Phys. Educ. Res.* 5, 013101-1 - 013101-5 (2009).

<sup>2</sup>R. S. Siegler and K. Crowley, "The microgenetic method: A direct means for studying cognitive development," *Am. Psych.* 46, 606-620 (1991).

<sup>3</sup>D. Kuhn, "Microgenetic study of change: what has it told us?," *Psych. Sci.* 6, 133-139 (1995).

<sup>4</sup>D. P. Maloney, T. L. O'Kuma, C. J. Hieggelke, C. J., and A. Van Heuvelen, "Surveying students' conceptual knowledge of electricity and magnetism," *Am. J. Phys.* 69, S12-S23 (2001).

<sup>5</sup>P. V. Engelhardt and R. J. Beichner, “Students' understanding of direct current resistive electrical circuits,” *Am. J. Phys.* 72, 98-115 (2004).

<sup>6</sup>G. B. Semb and J. A. Ellis, “Knowledge taught in school: What is remembered?,” *Rev. Educ. Res.* 64, 253-286 (1994).

<sup>7</sup>D. C. Rubin and A. E. Wenzel, “One hundred years of forgetting: A quantitative description of retention,” *Psychol. Rev.* 103, 734-760 (1996).

<sup>8</sup>S. M. Austin and K. E. Gilbert, “Student performance in a Keller-Plan course in introductory electricity and magnetism,” *Am. J. of Phys.* 41, 12-18 (1973).

<sup>9</sup>G. E. Francis, J. P. Adams, and E. J. Noonan, “Do they stay fixed?,” *Phys. Teach.* 36, 488- 480 (1998).

<sup>10</sup>A. Savinainen, P. Scott, and J. Viiri, “Using a bridging representation and social interactions to foster conceptual change: Designing and evaluating an instructional sequence for Newton’s third law,” *Sci. Ed.* 89, 175-195 (2005).

<sup>11</sup>R.K. Thornton, "Conceptual Dynamics: following changing student views of force and motion," in *AIP Conference Proceedings*, edited by E.F. Redish and J.S. Rigden (AIP, New York, 1997), **399**, 241-266.

<sup>12</sup>J. Petri and H. Niedderer, "A learning pathway in high school level quantum atomic physics", *Int. J. Sci. Ed.* **20**(9), 1075–1088 (1998).

<sup>13</sup>Reference 11.

<sup>14</sup>D. E. Meltzer, "How Do You Hit A Moving Target? Addressing The Dynamics Of Students' Thinking," *Proceedings of 2007 Physics Education Research Conference*. edited by. J. Marx, P. Heron, and S. Franklin. (Melville, New York: AIP Conference Proceedings, 2004), 7-10.

<sup>15</sup>L. Bao and E. F. Redish, "Model analysis: Representing and assessing the dynamics of student learning," *Phys. Rev. ST: Phys. Educ. Res.* **2**, 0101031 - 01010316 (2006).

<sup>16</sup>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**, 64-74 (1985).

<sup>17</sup>L. Ding, N. W. Reay, A. Lee, and L. Bao, “Effects of testing conditions on conceptual survey results,” *Phys. Rev. ST: Phys. Educ. Res.* 4, 0101121- 0101126 (2008).

<sup>18</sup>Reference 6.

<sup>19</sup>Reference 7.

<sup>20</sup>H. L. Roediger, III, “Relativity of remembering: Why the laws of memory vanished,” *Ann. Rev. of Psychol.* 59, 225-254 (2008)

<sup>21</sup>J. A. McGeoch, “Forgetting and the law of disuse,” *Psychol. Rev.* 39, 352–70 (1932).

<sup>22</sup>M. C. Anderson and J. H. Neely, “Interference and inhibition in memory retrieval,” in *Memory. Handbook of Perception and Cognition*, 2nd ed., edited by E. L. Bjork and R. A. Bjork (San Diego, CA: Academic Press, 1996), pp. 237-313.

<sup>23</sup>J. D. Karpicke and H. L. Roediger III, “The critical importance of retrieval for learning,” *Science* 319, 966-968 (2008).

<sup>24</sup>A. C. Butler and H. L. Roediger III, "Testing improves long-term retention in a simulated classroom setting," *Eur. J. Cog. Psychol.* 19, 514-527 (2007).

<sup>25</sup>C. Henderson, "Common concerns about the force concept inventory," *Phys. Teach.* 40, 542-547 (2002).

<sup>26</sup>J. C. Dutton, "WebAssign: A Better Homework Delivery Tool," *The Technology Source*, January/February (2001).

<sup>27</sup>D. Halliday, R. Resnick, and J. Walker, *Fundamentals of Physics*, 6<sup>th</sup> ed. (Wiley, New York, 2008).

<sup>28</sup>J. B. Willett, "Questions and answers in the measurement of change," in *Review of Research in Education*, Volume 15, edited by E.Z. Riothkopf. (American Educational Research Association, Washington DC, 1988).

<sup>29</sup>D. Rogosa, D. Brandt, and M. Zimowski, "A growth curve approach to the measurement of change," *Psychol. Bull.* 92, 726-748 (1982).

<sup>30</sup> As an analogy, one might consider that the theory of stellar evolution is empirically based on cross sectional data, since longitudinal data collection is not feasible, except for perhaps computer simulation ‘data’, the models of which must ultimately be based on cross sectional observations.

<sup>31</sup> Reference 1.

<sup>32</sup>I. J. Myung, “Tutorial on maximum likelihood estimation,” *J. of Math. Psychol.* 47, 90-100 (2003).

<sup>33</sup>I. A. Halloun and D. Hestenes, 1985 “The initial knowledge state of college physics students,” *Am. J. Phys.* 53 1044-1055 (1985).

<sup>34</sup>R. Rosenblatt, E. C. Sayre, and A. F. Heckler, “Toward a comprehensive picture of student understanding of force, velocity, and acceleration,” in *Proceedings of 2008 Physics Education Research Conference*. edited by C. Henderson, M. Sabella, L. Hsu. (AIP Conference Proceedings, Melville, New York, 2008).

<sup>35</sup> Reference 5.

<sup>36</sup> Reference 27.

<sup>37</sup>L. C. McDermott, and P. S. Shaffer, “Research as a guide for curriculum development: an example from introductory electricity. Part 1: Design of instructional strategies,” *Am. J. Phys.* 60, 994-1003 (1992).

<sup>38</sup>P. S. Shaffer and L. C. McDermott, “Research as a guide for curriculum development: And example from introductory electricity. Part 2: Investigation of student understanding,” *Am. J. Phys.* 60, 1003-1013 (1992).

<sup>39</sup> Reference 21.

<sup>40</sup> Reference 22.

<sup>41</sup>We are not claiming that the students are explicitly (and incorrectly) thinking “E is a scalar, so I have to answer in this way”, rather it is more likely that they are not thinking much at all about the need to distinguish the difference between vector and scalar quantities when determining E and V.

<sup>42</sup> Reference 1.

<sup>43</sup> There appears to be a dip in the direction-of-field response in weeks 3 and 4, and answering seems to be random during these weeks. It is not clear why this occurred.

For example there is no decrease in average student grade during this time. It may be due to some kind of interference as well.

<sup>44</sup> Reference 4.

<sup>45</sup>The interference of learning magnetic force on answering electric force questions was observed by T. Scaife and A. Heckler in a collaborative earlier work. A manuscript describing this in more detail is in preparation for publication.

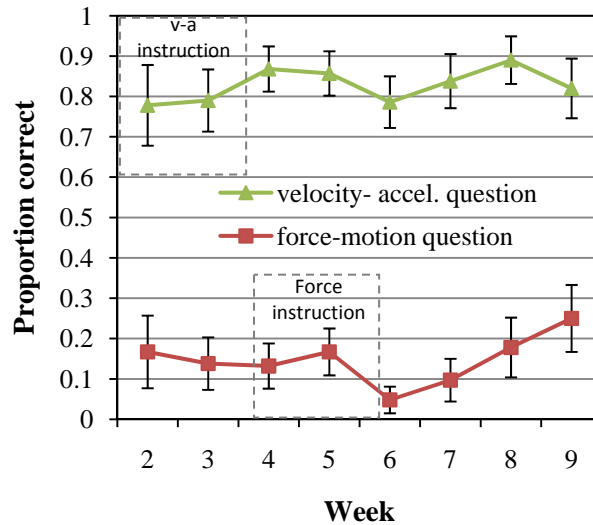
<sup>46</sup> Reference 1.

<sup>47</sup>F. N. Dempster, "Spacing effects and their implications for theory and practice," *Educ. Psychol. Rev.* 1, 309-330 (1989).

<sup>48</sup>R. Seabrooke, G. D. A. Brown, and J. E. Solity, "Distributed and massed practice: From laboratory to classroom," *App. Cog. Psychol.* 19, 107-122 (2005).

<sup>49</sup>P. I. Pavlic and J. R. Anderson, "Using a model to compute optimal schedule of practice," *J. Exp. Psychol. App.* 14, 101-117 (2008).

Figure 1



**Force-motion question:** At exactly 2:31PM, a boat is moving to the north on a lake. Which statement best describes the forces on the boat at this time?

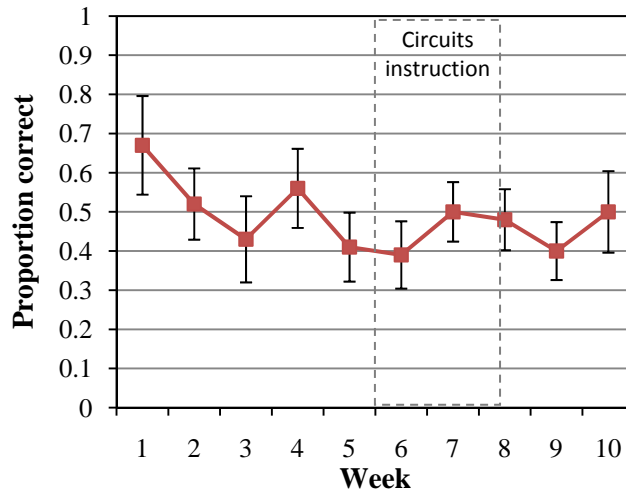
- There may be several forces to the north and to the south acting on the boat, but the forces to the north are larger.
- There may be several forces to the north and to the south acting on the boat, but the forces to the south are larger.
- There may be several forces to the north and to the south acting on the boat, but the forces to the south are equal in magnitude to those to the north.
- both a and b are possible.
- both a and c are possible
- a, b, and c are possible

**Velocity-acceleration question:** A car is moving to the right and speeding up. Which statement best describes the acceleration of the car?

- The car's acceleration is to the right.
- The car's acceleration is to the left.
- The car's acceleration equals zero.
- both a and b are possible
- both a and c are possible
- a, b, and c are possible

Figure 1. Example of flat response curves for two questions in mechanics, with an average of 32 students per week. Error bars represent 1 std.err. of mean.

Figure 2



Compare the brightness of the bulb in circuit 1 with that in circuit 2. Which bulb is BRIGHTER?

- The bulb in circuit 1 is brighter.
- The bulbs are the same brightness.
- The bulb in circuit 2 is brighter.

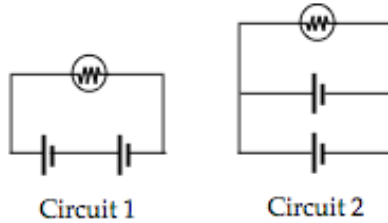
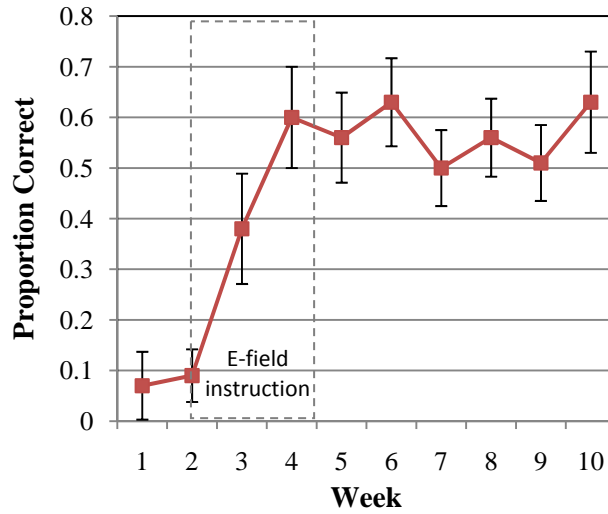


Figure 2. Example of a flat response curve for a DC circuit question, where all batteries are identical. Note no changes during instruction. Average of 32 students per week.

Error bars represent 1 std.err. of mean.

Figure 3



The image shows part of two infinite, flat plates, each with equal positive charge density evenly distributed over them.

What direction is the electric field at point C?

- a) To the left
- b) The field is zero
- c) To the right

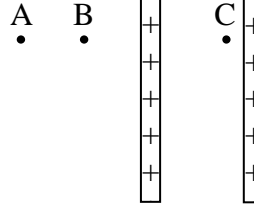


Figure 3. Example of a step-up response curve for an E-field question. Note increase during instruction, and no subsequent change. Average of 32 students per week. Error bars represent 1 std.err. of mean.

Figure 4

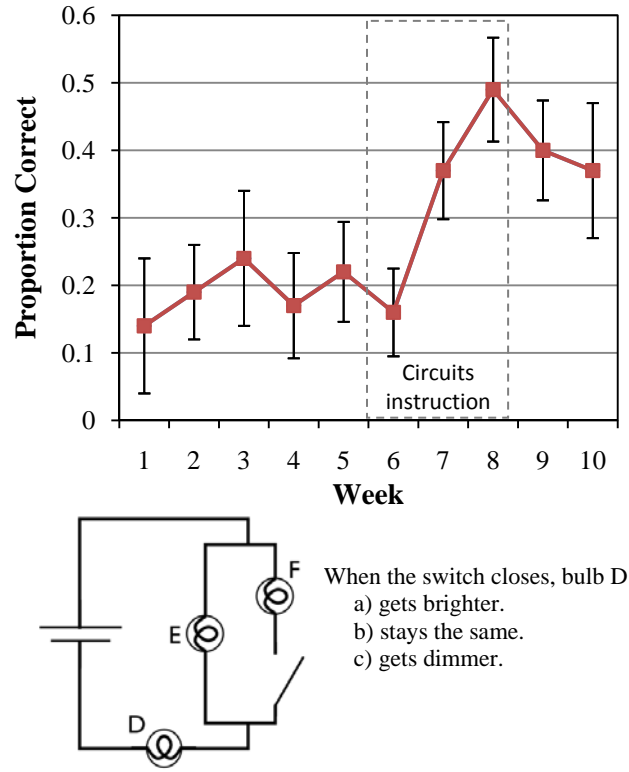
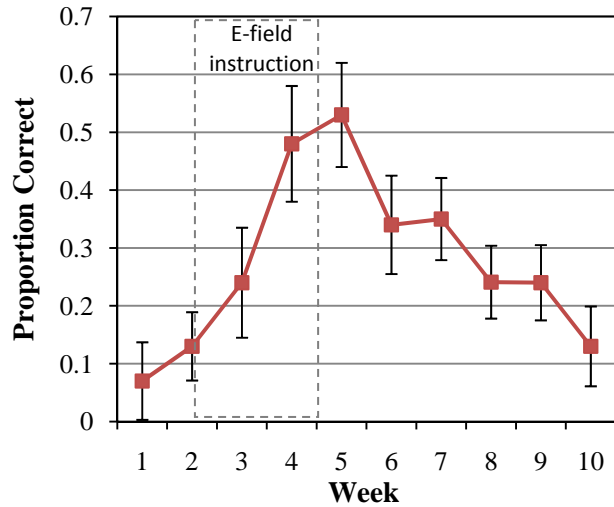


Figure 4. Example of a step-up response curve for a DC circuits question. Note increase during instruction, and no subsequent change. Average of 32 students per week. Error bars represent 1 std.err. of mean.

Figure 5



The image shows part of two infinite, flat plates, each with equal positive charge density evenly distributed over them.

Compare the magnitude of the electric field at points A and B.

- a)  $A > B$
- b)  $A = B$
- c)  $A < B$

A    B

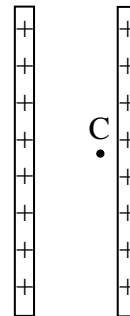


Figure 5. Example of a rise and decay response curve for an E-field question. Note increase during instruction, and subsequent decrease in performance. Average of 32 students per week. Error bars represent 1 std.err. of mean.

Figure 6

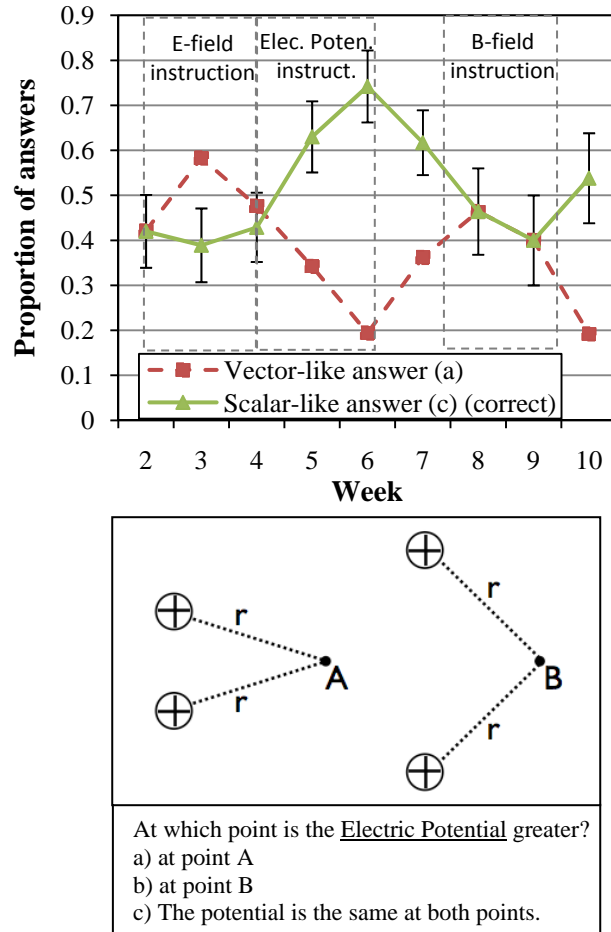


Figure 6. Example of a rise and decay response curve affected by interference for an electric potential question. Note the increase in correct, scalar-like answers during elec. potential instruction, and decrease during magnetic field instruction. In contrast, the vector-like answers tend to increase during instruction of vector fields (E and B). The increase in correct answers in the last week correspond to a review homework set (see Fig. 11). Average of 35 students per week. Error bars represent 1 std.err. of mean. Errors of vector-like answers are similar in size.

Figure 7

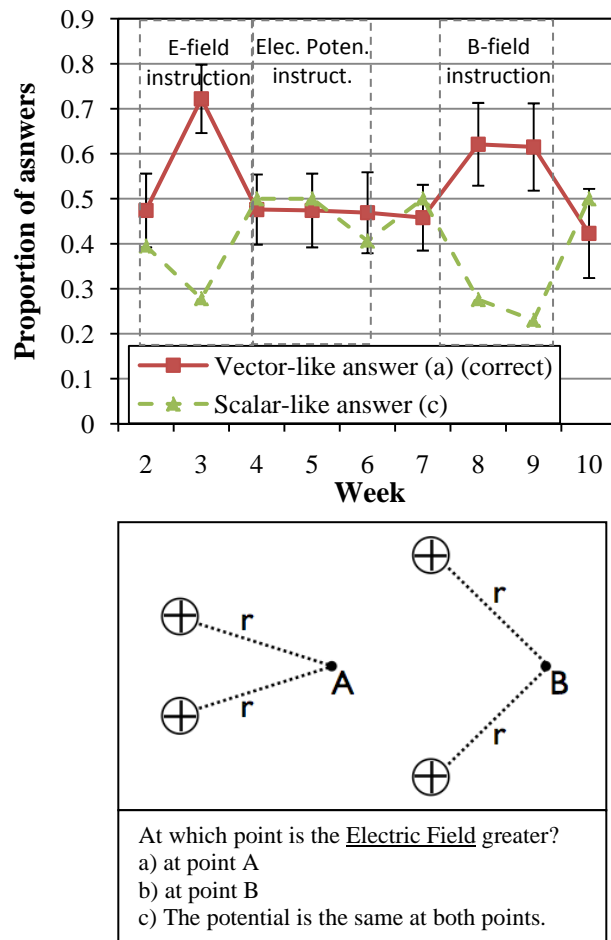
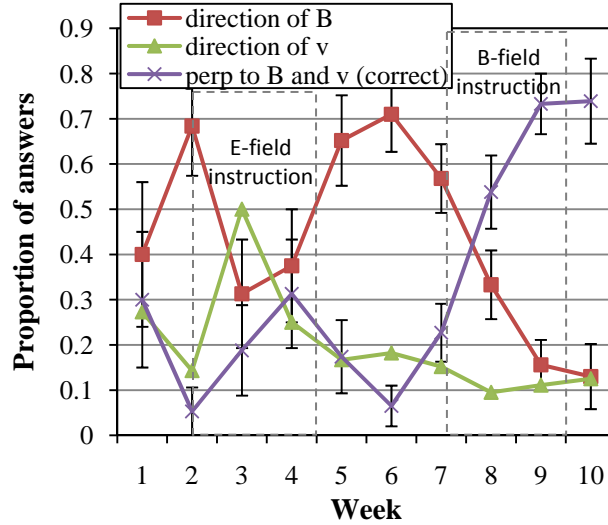


Figure 7. Example of a response curve affected by interference for an electric field question. Note the increase in correct, vector-like answers during E and B field (vector) instruction, and the decrease during electric potential instruction. In contrast, the scalar-like answers tend to increase during instruction of the scalar electric potential. Average of 35 students per week. Error bars represent 1 std.err. of mean. Errors of sc-like answers are similar in size.

Figure 8



Which of the following best describes the direction of the force experienced by the particle?

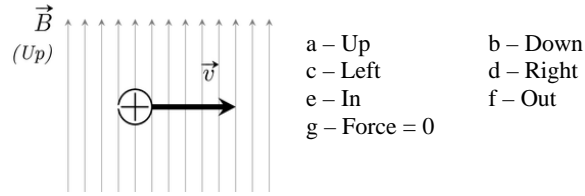
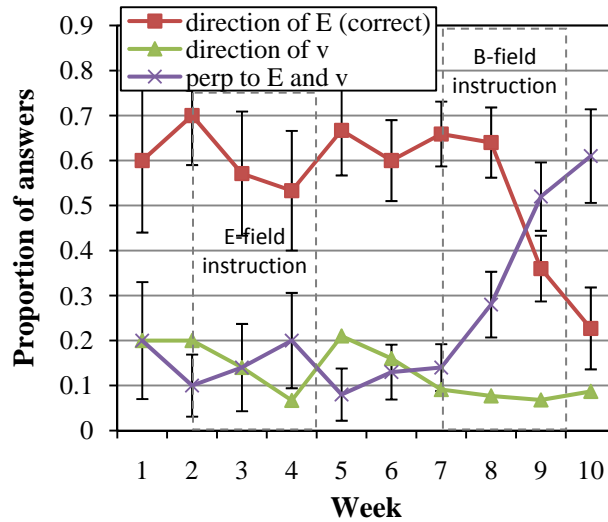
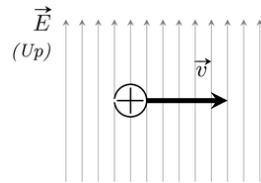


Figure 8. Example of a response curve affected by interference for a magnetic force question. Just after E-field instruction, the majority of answers in are in the direction of the B-field, similar to what they were taught for E-field and electric force. The correct answers rapidly increase during magnetic force instruction. It is not clear why the answers are somewhat random during E-field instruction. Average of 28 students per week. The three main responses shown. The “direction of B” response included both a and b responses. Error bars represent 1 std.err. of mean. Errors of direction-of-v answers are similar in size.

Figure 9



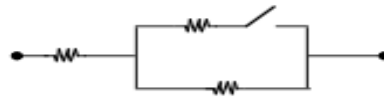
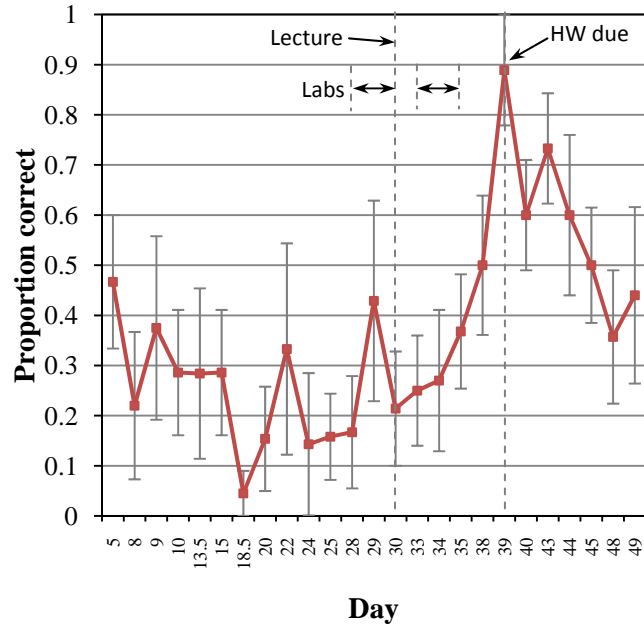
Which of the following best describes the direction of the force experienced by the particle?



- a – Up
- b – Down
- c – Left
- d – Right
- e – In
- f – Out
- g – Force = 0

Figure 9. Example of a response curve affected by interference for an electric force question. Note that before B-field instruction, the majority of answers are correct, in the direction of the electric field  $E$ . However, answers in the direction *perpendicular* to velocity  $v$  and Electric field  $E$  increase rapidly during magnetic force instruction, dramatically demonstrating that students are confusing electric force with magnetic force. Average of 28 students per week. The three main responses shown. The “direction of  $E$ ” response included both a and b responses. Error bars represent 1 std.err. of mean. Errors of direction-of- $v$  answers are similar in size.

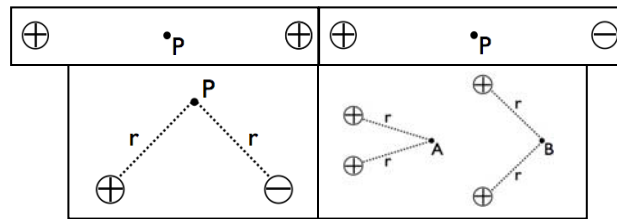
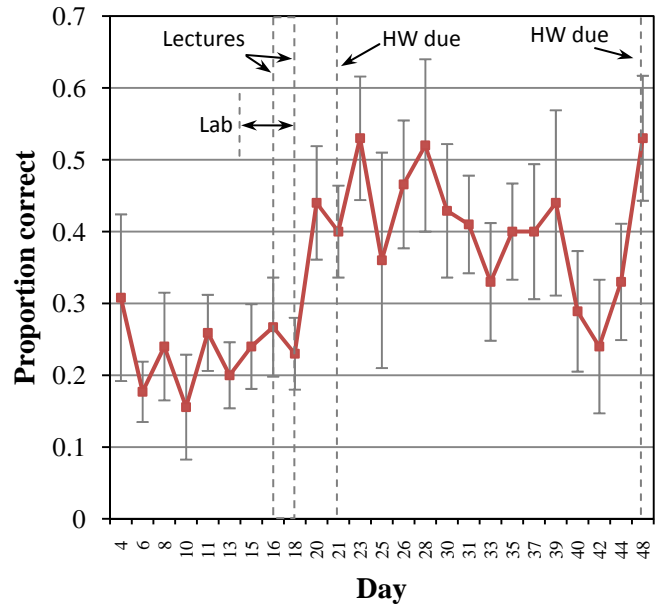
Figure 10



How does the resistance between the endpoints change when the switch is closed?  
 a) Increases  
 b) Stays the same  
 c) Decreases

Figure 10. Example of a high response curve for a DC circuits question, with a resolution on the order of one day. Note the rapid increase in score occurs one week after the relevant traditional lecture and labs, coinciding with the homework, which includes immediate feedback. Over 60% of student complete the homework with 2 days of due date. Average of 12 students per point. Days 13 & 14 were combined, as well as days 18 & 19 to reach the minimum of 6 students for each point. Note that the scale of the horizontal axis is not uniform. Error bars represent 1 std.err. of mean.

Figure 11



For the first three diagrams, student were asked to determine the electric potential at point P, with choices of up, down, right, left, positive, negative, or zero. For the last diagram students were ask to determine at which point, A or B, the electric potential was greater or if they had equal values.

Figure 11. Example of a high response curve for a collection of 4 electric potential questions, with a resolution on the order of one day. Note the rapid increase in score occurs two days after the relevant traditional lectures and labs (questions were answered *after* lectures on days 16 and 18), coinciding with the homework, which includes immediate feedback. Over 65% of student complete the homework with 2 days of due date. Note also the peak on day 48, when a relevant review homework was due. Average of 12 students per point, with a minimum 5 students per point. Note that the scale of the horizontal axis is not uniform. Error bars represent 1 std.err. of mean.