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

Andrew F. Heckler
Department of Physics, Ohio State University, Columbus, Ohio 43210

Eleanor C. Sayre
Department of Physics, Wabash College, Crawfordsville, Indiana 47933

(Received 2 September 2009; accepted 17 March 2010)

To characterize the evolution of student understanding better than what is possible by pre-and post-testing, we posed simple conceptual questions several times per week to separate, randomly selected groups of introductory physics students. This design avoids issues of retesting and allows for tracking of student understanding of a given topic during the course with a resolution on the order of days. Based on the data from over 1600 students over five quarters, we found significant and interesting changes in performance on time scales of days and weeks. We found that the response curves of performance versus time can be divided into four categories: flat (no change), step-up, step-up and decay, and step-down. We examined changes on a 1 day time scale and found that changes in performance did not coincide with relevant traditional lectures or laboratories, but 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. © *2010 American Association of Physics Teachers.*
[DOI: 10.1119/1.3384261]

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. Measuring changes in student understanding between pre- and post-tests can also be of significant interest. For example, measuring understanding throughout a course can help address questions such as does learning of a particular topic occur during lecture, homework, or some other parts 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 questions can be better answered by measuring student understanding on time scales of days and weeks throughout the course.

In this paper, we report on the evolution of student responses on several 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 these data in the form of "response curves," namely, graphs of performance on given questions as a function of time. In a recent paper,[1] we identified some interesting features such as peak-and-decay and temporary interference in a few response curves from a portion (representing one quarter) of the data presented here. Here, we report on a larger collection of data from ten classes ($\sim$1700 students, five quarters) to more systematically present and highlight results from the study.

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. We will operationally define a measurement of student understanding as the performance on simple conceptual multiple-choice problems. This definition has its limitations because 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[2,3] employ frequent quantitative and qualitative measures of complex learning processes such as problem solving. 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.

The validity of the simple multiple-choice questions used in this study was supported in several ways. After the questions were administered, we conducted informal, short postinterviews with most of the participants to verify that the questions were understood as intended and their answers were interpreted appropriately. To the extent that validity is partially supported from correlations with other measures of relevant knowledge, most of the questions used in this study were significantly correlated with the final course grade ($p < 0.05$). Most of the response curves show features that coincide with relevant instruction, lending further support to construct validity. A number of questions were derived from questions in more formally validated assessments such as CSEM (Ref. 4) and DIRECT,[5] or are questions commonly found in lectures or the conceptual question section in the end of the textbook chapters.

We have found no other studies that continually track understanding of populations of students on time scales of days or weeks during instruction. Most previous work on the evolution of student knowledge either studied the long-term retention of knowledge after the end of instruction and/or involved pre- and post-tests a few days or weeks before and after instruction. Outside of physics education, there have been a number of studies on long-term retention, typically on

 © 2010 American Association of Physics Teachers

the order of a few weeks to a few years after instruction, extending to up to 50 years, as summarized in extensive reviews.[6,7]

In the physics content area, there are a few small scale empirical studies of student retention[8–10] or short term changes in student understanding (see, for example, Refs. 11 and 12), and some discussion about models and methods of measuring changes in understanding.[11,13,14] 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 (see, for example, Ref. 15). Ding *et al.*[16] found that pre- and/or post-tests can change significantly by varying the day on which it is administered. They identified changes in particular questions relevant to the lectures just before the test as the cause, thus highlighting the importance of changes in understanding on smaller time scales.

There are several fundamental reasons to expect interesting changes in student performance on time scales of the order of days and weeks. Because forgetting of knowledge occurs on all time scales, with the largest loss occurring just after learning,[6,7,17] it is possible that much knowledge learned in a physics class may be forgotten within hours or days after learning. Also, although learning complex skills such as problem solving may take weeks or even years, learning simple concepts such as an electric field exists around a charged particle could happen quite rapidly. Finally, the learning of one topic may interfere with the understanding or learning another topic, as is well documented in Refs. 18 and 19. Because 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 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.[20,21] Although more complicated physics conceptual tests may have minimal retesting effects when the testing is separated by many weeks,[22] 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 quasirandomly selecting a group of students (that is, by randomly selected recitation sections) for testing each week of the course. Each group was considered to be statistically equivalent, allowing for comparison of performance between groups throughout the quarter.

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. To further verify that the results were not biased by groups with higher grades, we 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. The response curves reported here have obvious features (large effect sizes), and none of the curves of the average final course grade versus time or response curves of other test questions on unrelated topics follow the same pattern. In short, the response curves reported in this paper do not appear to be explainable by differences in the average grade between the groups or average performance on other unrelated questions. Thus, our assumption that the groups were equivalent prior to instruction is reasonable.

A. Participants and method

The participants in this study were 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. We studied two lecture sections each quarter for five quarters during the academic years 2007–8 and 2008–9. A total of 1694 students participated over five quarters, with roughly equal numbers of student each quarter. For each quarter, the instructor was a regular physics faculty member who taught both sections. There were four different faculties; three had 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 methodology (such as voting machines) used beyond the occasional answering of a question from a student. In addition to the three lectures per week, there was one recitation and a traditional laboratory with "cookbook-like" confirmatory laboratory activities.

Homework was assigned via WebAssign[23] once per week and consisted of $\approx$10 typical end of chapter problems.[24] 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 ten 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). This homework assignment consisted of participating in a 1 h session in our physics education research laboratory, where students completed 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 one to two laboratory sections (out of a total of about 14) and ask students to sign up for flexible homework. About 95% of the students typically participated in the flexible homework and all of these students participated in the study.

During the flexible homework session, students were asked to answer the questions as best they could even if they have not seen the material yet. 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. Afterward 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, laboratory book, and homework assignments to deter-

mine more precisely when relevant events in the course, such as a lecture or laboratory on a particular topic, occurred. Lectures were observed, and field notes on lecture content were recorded including the approximate level of attendance. We also collected the grades for each student in the course.

B. Analysis 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 2 week intervals, and perform a chi-squared test of independence for each binning. This phase 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 was used over the course of the study, and we found that there was significant variation in performance on 37 (about 30%) of the items. Binning did not qualitatively change this result. For all but two of the questions, the patterns were replicated in more than one quarter. The two that were not replicated (the questions in Sec. V and VI, Figs. 3 and 5) were administered in only one quarter, and the results were very large and significant. Because 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, the performance on specific questions between classes with different instructors was largely similar, and the method used here could be used to design more focused studies comparing instructional methods.

The second phase of the data analysis consisted of an examination of the response curves of the questions with significant variation to determine whether the significant variations occurred during relevant instructional events and to determine the general pattern of the variation. 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. In this phase we noticed four general categories of response curves (flat, step-up, rise and decay, and step-down), and these categories coincide well with the general causal mechanisms of learning, decay, and interference.

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 easily distinguishable, many of them could also be verified as statistically reliable by inspecting the response curves including error bars.

Although 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 Ref. 1. Such a quantitative analysis, for example, using maximum likelihood estimation methods,[25] allows for more precise quantitative parametrization, comparisons, and hypotheses testing. Such an analysis is beyond the scope of this paper and is a fertile area for more in-depth investigation.

C. Comment on nonlongitudinal data

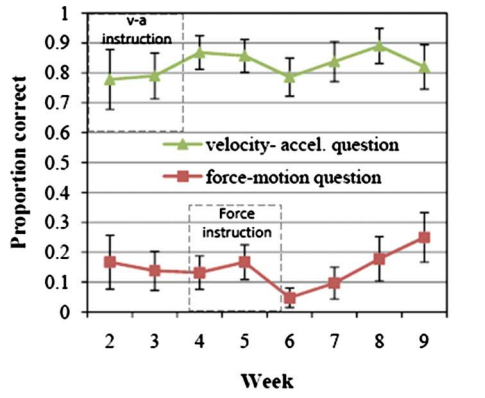
One must be careful in making inferences about the evolution of individuals from nonlongitudinal data. In particular, we must keep in mind that the nature of the response curves represent the evolution of the average of the population and does not necessarily imply that individuals follow this same path. For example, a "rise and decay" population average response curve could be comprised of two subpopulations, 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.[26,27]

Nonetheless, from a larger perspective, cross sectional data, such as the evolution of the class average, can produce useful information to instructors and education researchers, and some conclusions can still be made about the evolution of individuals. For example, we can rule out the hypothesis that a significant population of students individually follows a given pattern if the response curve does not follow this pattern. We can also make useful and reasonable assumptions about factors that may separate subpopulations of students into different response curves, such as math ability or final grade, and increase our confidence that a significant number of students in a given subpopulation are following or not following a given path.[28]

IV. FLAT CURVES: PERFORMANCE AT CEILING OR FLOOR?

As stated, about 70% of the conceptual questions resulted in no significant variation in student performance over the quarter. This result is consistent with several seminal physics education research studies, which found a lack of change from pre- to post-test for many simple conceptual questions (see, for example, Refs. 29 and 15). 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 is striking to see smooth, flat curves, that is, no change on a given question over the course of the quarter, even during instruction relevant to the question.

Figures 1 and 2 present examples of flat response curves and the corresponding questions. For these and most of the other graphs in the paper, the time window of the relevant instruction, which includes lecture, homework, and laboratory, 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.[30] The first question in Fig. 1 tests for the 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 an easy one 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.[5] The scores on this question do not change over the quarter ($\chi^2(9)=6.0,\ p=0.74$).



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?

a) There may be several forces to the north and to the south acting on the boat, but the forces to the north are larger.
b) There may be several forces to the north and to the south acting on the boat, but the forces to the south are larger.
c) 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.
d) both a and b are possible.
e) both a and c are possible
f) 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?

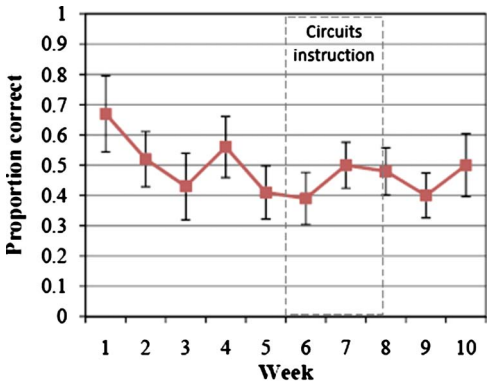
a) The car's acceleration is to the right.
b) The car's acceleration is to the left.
c) The car's acceleration equals zero.
d) both a and b are possible
e) both a and c are possible
f) a, b, and c are possible

Fig. 1. Example of flat response curves for two questions in mechanics, with an average of 32 students per week. Error bars in all the figures represent 1 standard deviation of the mean.

The upper curve in Fig. 1 with an average score of 83% is likely at ceiling (that is, effectively at maximum performance), and the lower curve with an average of 14% is likely at floor (that is, effectively at minimum performance). It might also be the case that the question in Fig. 2, with an average score of 47%, is also scoring near floor. Because there are three possible answer choices for this question and only 15% of students chose "equal," most students choose between "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, although from our informal debriefings with students, this possibility seems highly unlikely. The result that flat curves were either at floor or ceiling was consistent throughout our study.

V. STEP-UP

The "step-up" in performance response curve is characterized by an initial interval of no change in performance, followed by a rapid increase to a new level of performance that is maintained for the remainder of the course. We found step-up curves in about 15% of our items. All step-up curves 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. Figure 3



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

a) The bulb in circuit 1 is brighter.
b) The bulbs are the same brightness.
c) The bulb in circuit 2 is brighter.

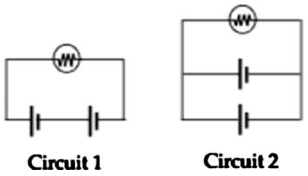
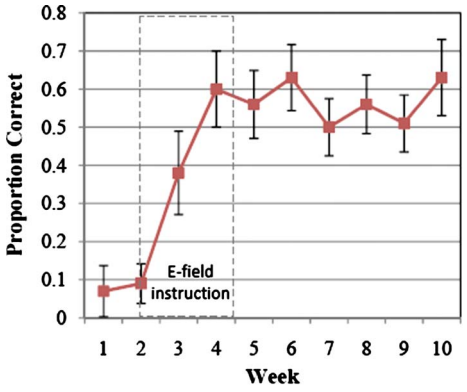


Fig. 2. Example of a flat response curve for a DC circuit question; all batteries are identical. There are no changes during instruction. Average of 32 students per week.

presents the response curve for a question that might be found in a typical lecture on the electric field or in the conceptual questions section at the back of a textbook chapter.[24] The weekly average proportion of correct responses changes over the course, $\chi^2(9)=38.6$, $p<0.0001$, with a clear step-up response showing a significant difference between the average scores before (9%) and after (54%) instruction, $\chi^2(1)=33.9$, $p<0.0001$, and Cohen's effect size of $d=1.1$.[31] The



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

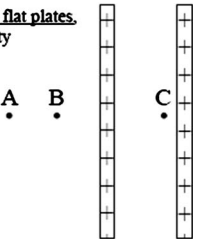


Fig. 3. Example of a step-up response curve for an *E*-field question. There is an increase during instruction and no subsequent change. Average of 32 students per week.

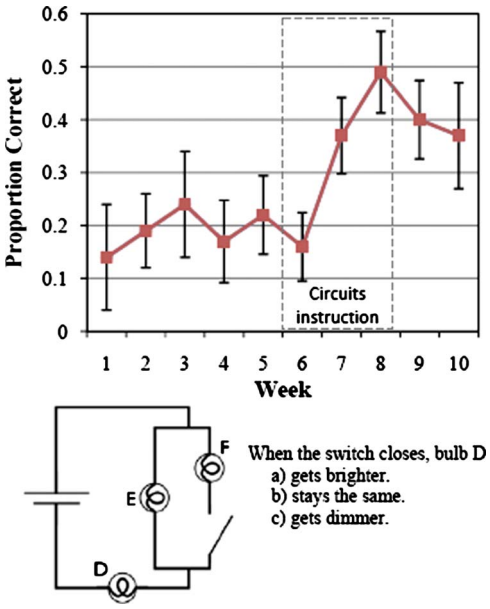


Fig. 4. Example of a step-up response curve for a DC circuit question. Average of 32 students per week.

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 that students have with understanding circuits.[32,33] 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, $\chi^2(1)=16.1$, $p<0.0001$, and $d=0.5$.

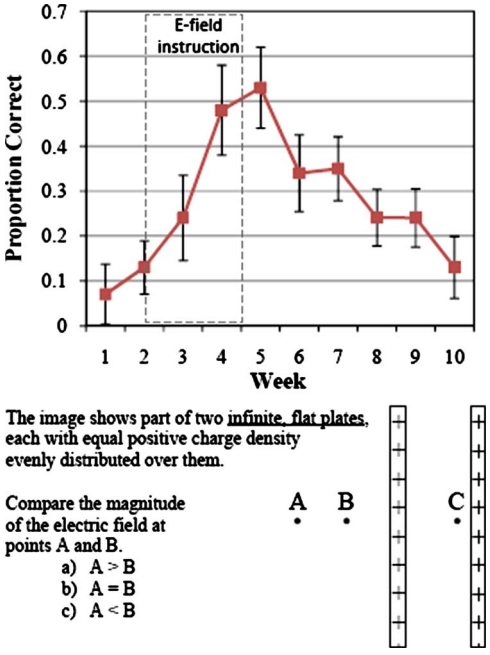


Fig. 5. Example of a rise and decay response curve for an *E*-field question. There is an increase during instruction and subsequent decrease in performance. Average of 32 students per week.

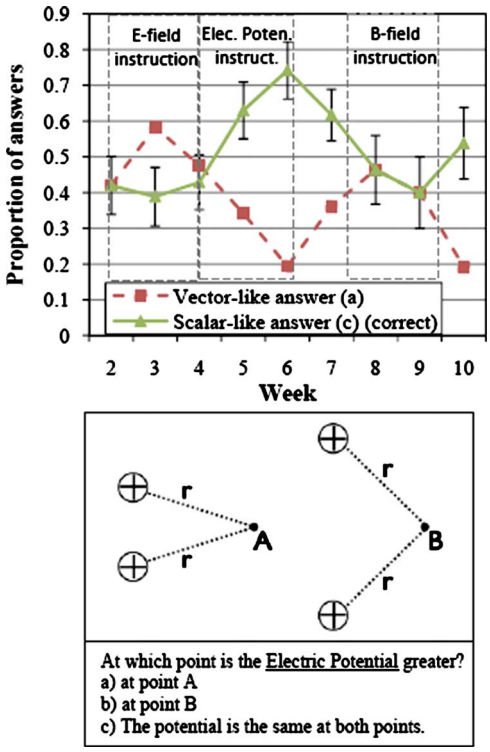


Fig. 6. Example of a rise and decay response affected by interference for an electric potential question. There is an increase in correct, scalarlike answers during electric potential instruction and a decrease during magnetic field instruction. The vectorlike answers increase during instruction of vector fields (E and B). The increase in correct answers in the last week corresponds to a review homework set (see Fig. 11). Average of 35 students per week. Errors of vectorlike answers are similar in size.

VI. RISE-DECAY ON SCALE OF WEEKS

To the extent that a particular topic is addressed for only a limited time during the course, it might be expected that student performance would rise during instruction and decay afterward. We found about 10% of our items follow this pattern. Figure 5 shows 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 pattern might be an example of a misconception returning after instruction.

Figure 6 presents a rise and decay curve, though the curve is more complicated. There is another rise in the last week, which is likely due to a review homework assignment at the end of the quarter that included similar questions. This particular rise will be discussed more in Sec. VIII. Although the performance changes significantly during the course, $\chi^2(8)=16.2$, $p=0.04$, there is reason to believe that the decrease in performance is due to explicit interference with a related topic.

VII. INTERFERENCE

When two topics are perceived as related, the learning of one may interfere with the learning and memory of the other.

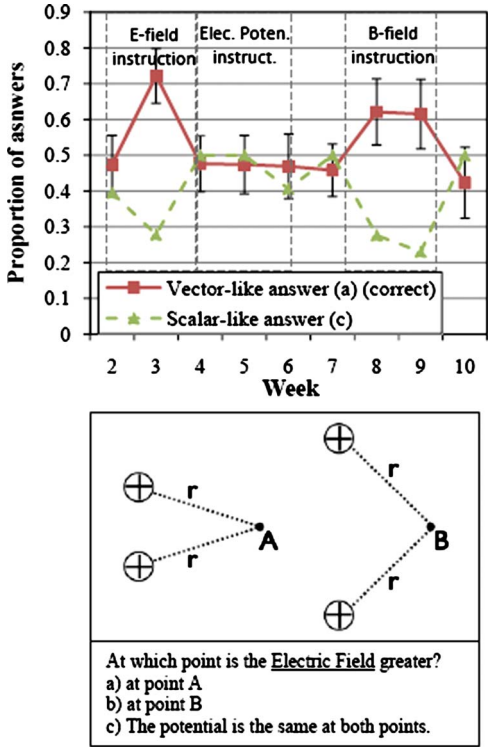


Fig. 7. Example of a response curve affected by interference for an electric field question. There is an increase in correct, vectorlike answers during *E* and *B* field (vector) instruction, and a decrease during electric potential instruction. The scalarlike answers tend to increase during instruction of the scalar electric potential. Average of 35 students per week. Errors of scalarlike answers are similar in size.

For example, much evidence suggests that a major cause of forgetting a particular piece of knowledge is the learning of new knowledge, especially if the new knowledge is in some way similar.[17,18] Therefore, the decay (forgetting) curves from Sec. VI could be viewed as a special case of interference, where the interference is small and continuous. In this section we will consider interference when the decrease in performance is large and rapid (order of days) and is coincident with the introduction of a similar topic that would plausibly give rise to interference. We found about 5% of our items exhibited interference.

The first example of interference involves vector and scalar quantities in electromagnetism. We have found that students often fail to recognize the importance of remembering the vector nature of the electric field *E* and the scalar nature of the electric potential *V*. This confusion might occur because the concepts of electric field and electric potential are both unfamiliar, abstract quantities used in electrostatics and are perceived as two highly similar quantities by students. As a result, *E* and *V* may interfere with each other in the course of learning and problem solving. 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.[34]

In Ref. 1 we provided evidence that learning about the electric potential interferes with student understanding of the vector nature of an electric field. The example of interference in Fig. 7 in this paper replicates this finding and extends it to demonstrate that not only does learning about electric potential interfere with the understanding of electric field, but learning about the electric field (and magnetic field) can interfere with understanding of electric potential. In particular, as shown in Fig. 7, students correct "vectorlike" responses to *E*-field questions increased during instruction about *E*, then quickly turned to the scalarlike responses during the instruction on *V*, and then returned to vectorlike 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%), $\chi^2(1)=5.8$, $p=0.02$, and $d=0.3$. Conversely, as shown in Fig. 6, during electric field instruction, students begin to use (incorrect) vectorlike responses, then turn to (correct) scalarlike responses during electric potential instruction, and then return to vectorlike 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%), $\chi^2(1)=7.5$, $p<0.01$, and $d=0.3$.

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

The second example of interference involves the electric force and the magnetic force. Because the representations of vector *E* fields and *B* fields are often similar, and both are invisible fields that exert a force on a charged particle, we might expect that students can 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.[35] Following this phase, there was a clear rise in the correct responses during/after magnetic force instruction (63%) compared to before instruction (14%), $\chi^2(1)=66.1$, $p<0.0001$, and $d=1.1$. Consistent with Ref. 4, this result 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, because they have not been taught otherwise. Once taught, many students learn magnetic force quickly.

The more dramatic signal of interference[36] 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. Once magnetic force was taught, there was a sudden decrease to 44% of correct answers in the direction of the field and a 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*, $\chi^2(1)=34.8$, $p<0.0001$, and $d=0.8$. By comparing Figs. 8 and 9, we see that movement away from the correct electric force answer occurs only about 1–2 weeks after there is significant learning of the *B*-field force.

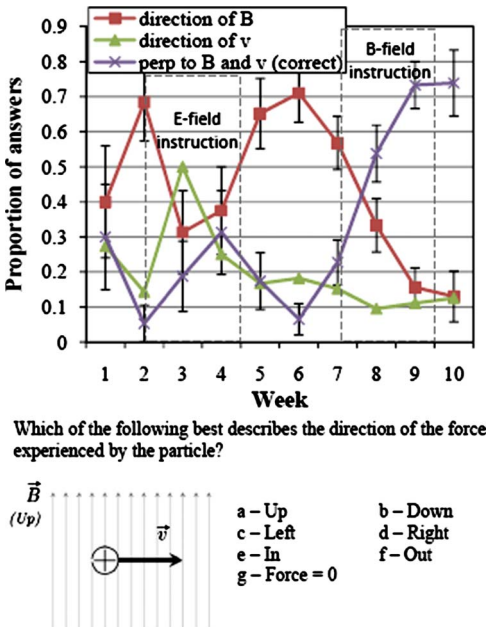


Fig. 8. Example of a response curve affected by interference for a magnetic force question. Just after *E*-field instruction, most answers are in the direction of the *B*-field, similar to what they were taught for *E*-field and electric force. The correct answers rapidly increase during magnetic force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

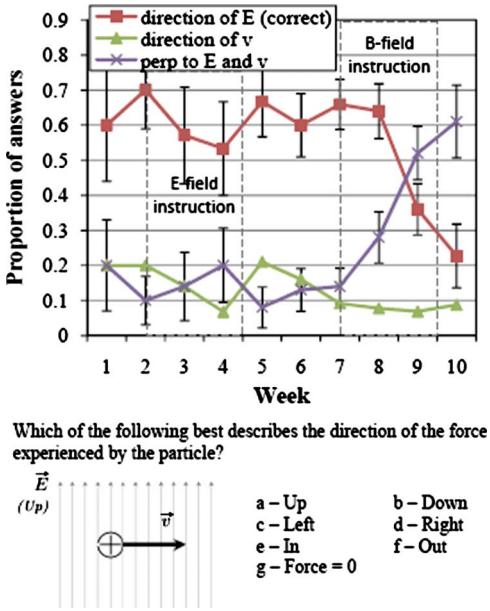


Fig. 9. Example of a response curve affected by interference for an electric force question. Before *B*-field instruction, the majority of answers are correct, in the direction of the electric field *E*. Answers in the direction perpendicular to the velocity *v* and electric field *E* increase rapidly during magnetic force instruction, demonstrating that students confuse electric force with magnetic force. The three main responses are shown. The "direction of *E*" response includes both a and b. Response errors of direction-of-*v* answers are similar in size.

VIII. RAPID CHANGES

We have sections discussed changes in student performance on a time scale of weeks. Because data were typically collected several times per week, we can observe changes on the time scale of days. This increased time resolution can help us answer some 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, laboratory, or recitation.

The example shown in Fig. 10 is a question about circuits that is closely related to a question in DIRECT.[5] There are two notable features of this response curve. First, there is a rapid rise in performance, which does not coincide with the relevant lecture or laboratories, but coincides with a relevant homework set. In particular, the rapid rise begins on quarter day 38 and the relevant lecture occurred on day 30 (attendance is ≈60%) in which an explicit example was presented with the same combination of resistors and a closely related demonstration. There were two weeks of laboratories on circuits including multiple loop circuits, on days 28–30 and 33–35. The relevant homework assignment was due on day 39 and included three problems on multiple loop circuits, including one very similar to the question in Fig. 10. Because the homework was online, we looked at electronic records of students' homework activity and found that over 60% of the students completed the relevant problems within 2 days of the deadline. By 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, $\chi^2(1)=29.6$, $p<0.001$, and $d=0.86$. Thus, it is likely that the

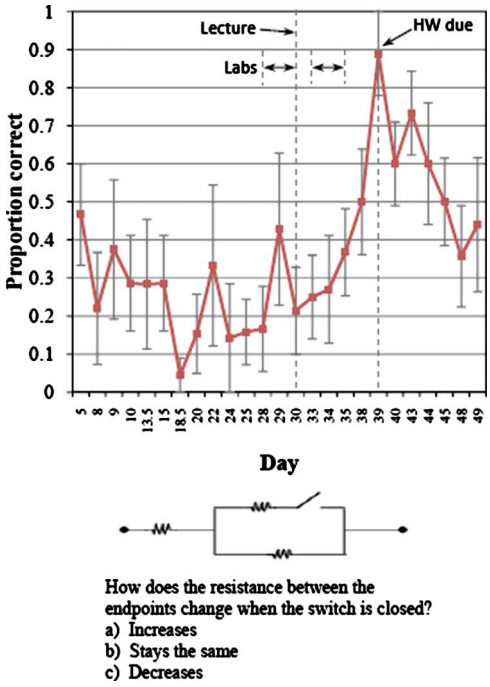


Fig. 10. Example of a high response curve for a DC circuit question, with a resolution on the order of 1 day. Note the rapid increase in score occurs 1 week after the relevant traditional lecture and laboratories, coinciding with the homework, which includes immediate feedback. Over 60% of student complete the homework with 2 days of due date. Average of 12 students per point. Days 13 and 14 were combined, as well as days 18 and 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 standard error of the mean.

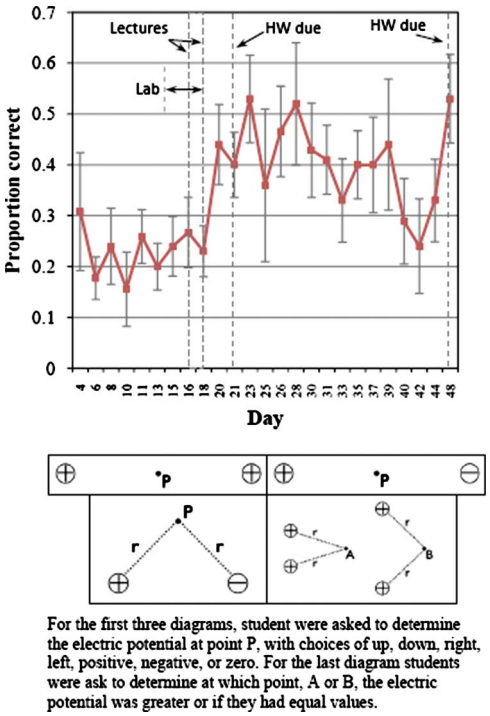


Fig. 11. Example of a high response curve for a collection of four electric potential questions, with a resolution on the order of 1 day. Note the rapid increase in score occurs 2 days after the relevant traditional lectures and laboratories (questions were answered *after* lectures on days 16 and 18), 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 standard deviation of the mean.

homework is the cause of the large and rapid rise in performance and not traditional lecture or laboratory.

The second notable feature of the response curve in Fig. 10 is the decrease in performance over the time of several days to 1 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 1 week after homework, $\chi^2(1)=3.8$, $p=0.05$, and $d=0.37$. This finding is consistent with the result in Ref. 1, suggesting that the student performance can decrease significantly over even 1 day. A more precise parametrization of the decay time requires more careful modeling.

The example in Fig. 11 yields a response curve for the average of a collection of four similar questions about the electric potential resulting from two point charges of the same 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 laboratory. In particular, there is a distinct rise in performance on day 20, while the relevant lectures (attendance is ~70%) occurred on days 16 and 18, the relevant laboratories occurred on days 14–18, and the relevant homework was due on day 21. More than 65% of the students completed the relevant homework problems less than 2 days before the homework was due. In particular, by comparing the performance before the relevant homework, from days 1–19 to the performance for days 20–25 (during and just after when homework was due), there was a significant difference in performance, $\chi^2(1)=30.5$, $p<0.001$, and $d=0.8$. Two other features are worth noting. One is the dip in performance around day 42, which is likely due to an interference effect because the vectorlike responses peaked at this time. There is an increase in performance on day 48, which coincides with a review homework assignment in which a question about the electric potential from two charges is asked. Thus, it appears that this increase in performance is due to the review homework.

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

IX. CONCLUDING REMARKS

We found significant changes in student scores on simple conceptual questions on both the day and week time scales over the duration of a course. We represented the results as response curves of average score verses time and found four basic patterns with associated simple causes for each pattern.

The flat pattern indicates that there is 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 pattern can be simply explained by the possibility that the students were significantly above ceiling or below floor. Another simple explanation is that the instruction itself is ineffective, and the existence of a flat response only at ceiling or floor is a coincidence. Whether the flat responses are due to the nature of the question or instruction or both remains open.

For the step-up pattern, there is 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 is unexpected for there to be an (apparently) unchanging plateau after instruction. This plateau might be due to a long decay time, or to the possibility that the topic is constantly practiced at some level; the plateau presumably represents a ceiling in performance.

Another response pattern is a relatively rapid increase, followed by a significant decrease. This peak can happen on the time scale of days or weeks. This response highlights the fact that student performance can change both up and down rapidly over a range of time scales. The increases shown in this study were coincident with instruction, and the decreases may be due to forgetting, explicit interference from a related topic or some other factors. 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 type of response is a rapid decrease in score that coincides with the learning of a related topic. For the examples we have discussed, we argue that this type of response is likely due to interference.

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

First, the potential for rapid changes in scores, especially 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 can rapidly peak for a particular topic, even if only momentarily, instructors might be able to adjust instruction to extend the peaks. For example, there are several studies indicating that repeated and appropriately spaced practice can dramatically increase retention.[37–39]

Third, 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 their instruction to address this issue. Evidence of interference is an example of the power of high resolution information. Accurately measuring when the interference occurs aids in the identification of the specific interfering topic. Examining the dynamics of student models of understanding is a rich topic for further study.

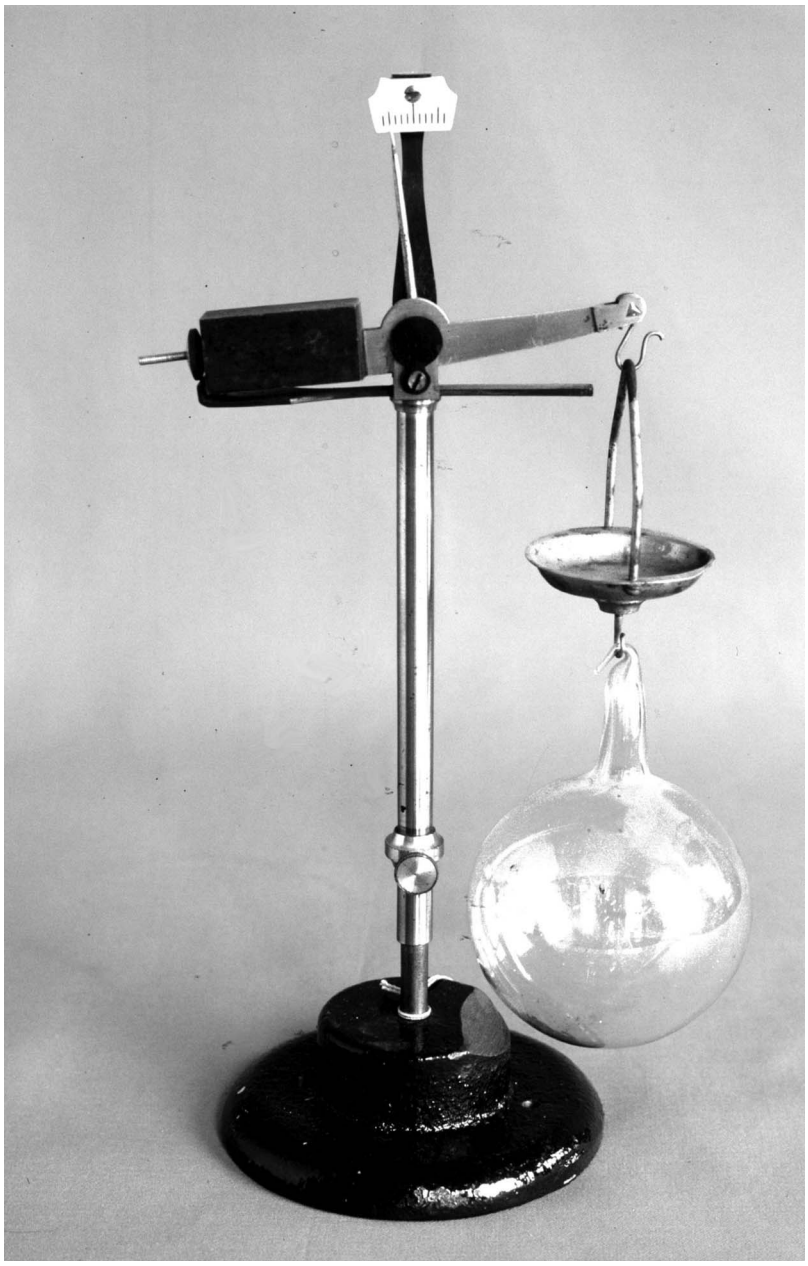
Finally, it can be very useful to know when student understanding improves during the course, and what instructional events caused this understanding to occur. In this paper, the increases in performance happened only directly after homework (with feedback) and not directly after traditional lecture or laboratory. This difference is another reason to reconsider the value of traditional lectures and laboratories. 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 was partially supported by a grant from the Institute of Education Sciences, U.S. Department of Education (Grant No. R305H050125). The authors would like to thank Thomas Scaife and Rebecca Rosenblatt for help in designing questions and collecting data.

[1] 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 (2009).
[2] R. S. Siegler and K. Crowley, "The microgenetic method: A direct means for studying cognitive development," Am. Psychol. **46**, 606–620 (1991).
[3] D. Kuhn, "Microgenetic study of change: What has it told us?," Psychol. Sci. **6**, 133–139 (1995).
[4] D. P. Maloney, T. L. O'Kuma, C. J. Hieggelke, and A. Van Heuvelen, "Surveying students' conceptual knowledge of electricity and magnetism," Am. J. Phys. **69**, S12–S23 (2001).
[5] P. V. Engelhardt and R. J. Beichner, "Students' understanding of direct current resistive electrical circuits," Am. J. Phys. **72**, 98–115 (2004).
[6] G. B. Semb and J. A. Ellis, "Knowledge taught in school: What is remembered?," Rev. Educ. Res. **64**, 253–286 (1994).
[7] D. C. Rubin and A. E. Wenzel, "One hundred years of forgetting: A quantitative description of retention," Psychol. Rev. **103**, 734–760 (1996).
[8] S. M. Austin and K. E. Gilbert, "Student performance in a Keller-Plan course in introductory electricity and magnetism," Am. J. Phys. **41**, 12–18 (1973).
[9] G. E. Francis, J. P. Adams, and E. J. Noonan, "Do they stay fixed?," Phys. Teach. **36**, 488–490 (1998).
[10] A. Savinainen, P. Scott, and J. Viiri, "Using a bridging representation and social interactions to foster conceptual change: Designing and evaluating and instructional sequence for Newton's third law," Sci. Educ. **89**, 175–195 (2005).
[11] R. K. Thornton, "Conceptual dynamics: Following changing student views of force and motion," AIP Conf. Proc. **399**, 241–266 (1997).
[12] J. Petri and H. Niedderer, "A learning pathway in high school level quantum atomic physics," Int. J. Sci. Educ. **20** (9), 1075–1088 (1998).
[13] D. E. Meltzer, "How do you hit a moving target? Addressing the dynamics of students' thinking," in *Proceedings of 2007 Physics Education Research Conference*, edited by J. Marx, P. Heron, and S. Franklin (AIP, Melville, NY, 2004), pp. 7–10.
[14] L. Bao and E. F. Redish, "Model analysis: Representing and assessing the dynamics of student learning," Phys. Rev. ST Phys. Educ. Res. **2**, 010103 (2006).
[15] 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 (1998).
[16] 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**, 010112 (2008).
[17] H. L. Roediger III, "Relativity of remembering: Why the laws of memory vanished," Annu. Rev. Psychol. **59**, 225–254 (2008).
[18] J. A. McGeoch, "Forgetting and the law of disuse," Psychol. Rev. **39**, 352–370 (1932).
[19] 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 (Academic, San Diego, CA, 1996), pp. 237–313.
[20] J. D. Karpicke and H. L. Roediger III, "The critical importance of retrieval for learning," Science **319**, 966–968 (2008).
[21] A. C. Butler and H. L. Roediger III, "Testing improves long-term retention in a simulated classroom setting," Eur. J. Cogn. Psychol. **19**, 514–527 (2007).
[22] C. Henderson, "Common concerns about the force concept inventory," Phys. Teach. **40**, 542–547 (2002).
[23] J. C. Dutton, "WebAssign: A better homework delivery tool," Technology Source, January/February 2001.
[24] D. Halliday, R. Resnick, and J. Walker, *Fundamentals of Physics*, 6th ed. (Wiley, New York, 2008).
[25] I. J. Myung, "Tutorial on maximum likelihood estimation," J. Math. Psychol. **47**, 90–100 (2003).
[26] J. B. Willett, "Questions and answers in the measurement of change," in *Review of Research in Education*, edited by E. Z. Riothkopf (American Educational Research Association, Washington, D.C., 1988), Vol. 15.
[27] D. Rogosa, D. Brandt, and M. Zimowski, "A growth curve approach to the measurement of change," Psychol. Bull. **92**, 726–748 (1982).
[28] As an analogy, we might consider that the theory of stellar evolution is empirically based on the cross sectional data because longitudinal data collection is not feasible, except for perhaps computer simulation data, the models of which must ultimately be based on the cross sectional observations.
[29] I. A. Halloun and D. Hestenes, "The initial knowledge state of college physics students," Am. J. Phys. **53**, 1043–1055 (1985).
[30] 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, and L. Hsu (AIP, Melville, NY, 2008).
[31] Cohen's effect size d is a standard measure that quantifies the size of the difference between the means of two samples. It is defined as the number of standard deviations one mean is from the other: $d=(\text{mean}_1-\text{mean}_2)/s$, where s is the pooled standard deviation of the two samples.
[32] 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).
[33] 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).
[34] 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.
[35] There appears to be a dip in the direction-of-field response in weeks 3 and 4, and answers seem to be random during these weeks. It is not clear why this response occurred. For example, there is no decrease in average student grade during this time. It might be due to some kind of interference as well.
[36] The interference of learning magnetic force on answering electric force questions was observed by T. Scaife and A. Heckler in earlier work. A

manuscript describing this observation is in preparation.
[37]F. N. Dempster, "Spacing effects and their implications for theory and practice," Educ. Psychol. Rev. **1**, 309–330 (1989).
[38]R. Seabrook, G. D. A. Brown, and J. E. Solity, "Distributed and massed practice: From laboratory to classroom," Appl. Cognit. Psychol. **19**, 107–122 (2005).
[39]P. I. Pavlik and J. R. Anderson, "Using a model to compute optimal schedule of practice," J. Exp. Psychol., Appl. **14**, 101–117 (2008).



Buoyancy Balance. Chemists, who do careful weighing, know that we live at the bottom of a sea of air, and that a buoyant force equal to the weight of the air displaced by a body acts upward on it. Because the density of air is small, the buoyant force is also small. To demonstrate this effect, a glass balloon is placed on one end of a small equal arm balance placed under a bell jar. When the air surrounding the balloon is pumped out, the buoyant force is removed, and the balance tilts. This demonstration of the buoyant effects of the atmosphere, ca. 1900, at Denison University in Granville, Ohio is unmarked. (Photograph and Notes by Thomas B. Greenslade, Jr., Kenyon College)