Learning, retention, and forgetting of Newton's third law throughout university physics

Eleanor C. Sayre
Department of Physics, Kansas State University, Manhattan, Kansas 66506, USA

Scott V. Franklin and Stephanie Dymek
Department of Physics, Rochester Institute of Technology, Rochester, New York 14623, USA

Jessica Clark
Department of Physics and Astronomy, University of Maine, Orono, Maine 04469, USA

Yifei Sun
Department of Physics, Wabash College, Crawfordsville, Indiana 47933, USA
(Received 20 April 2011; revised manuscript received 9 March 2012; published 10 April 2012)

We present data from a between-student study on student response to questions on Newton's third law given in two introductory calculus-based physics classes (Mechanics and Electromagnetism) at a large northeastern university. Construction of a response curve reveals subtle dynamics in student learning not capturable by pretesting and post-testing. We find a significant positive effect of instruction that diminishes by the end of the quarter. Two quarters later, a significant dip in correct response occurs when instruction changes from the vector quantities of electric forces and fields to the scalar quantity of electric potential. When instruction returns to vector topics, performance rebounds to initial values.

DOI: 10.1103/PhysRevSTPER.8.010116 PACS numbers: 01.40.Fk, 01.30.lb

I. INTRODUCTION

Research on forgetting and interference [1–4] shows that learning is very subtle, and often time dependent, with even significant gains sometimes short lived. Sayre and Heckler have applied the between-student "response curves method" (RCM) [5,6] to physics classes. In the RCM, data are collected regularly through the academic term with comparison between different groups of students. While this requires significantly larger student populations—and overhead on grouping—when successful it allows for a much more detailed picture of student understanding before, during, and after instruction.

Pretesting and post-testing of students is virtually the standard for assessing learning in large physics classes [7]. Thornton and Sokoloff [8] used the method to establish the validity of the Force and Motion Concept Evaluation and to demonstrate the efficacy of active engagement classrooms, a study reproduced on a much larger scale by Hake [9]. Pretesting and post-testing fails, however, to reveal the rich dynamism of student learning. In physics education research, researchers often turn to interviews and case studies to investigate this dynamism. While these methods gather rich data and make robust observations, they are resource intensive and (consequently) follow smaller numbers of students. In the RCM, we balance resource constraints against a need to study the dynamism of student understanding in large enrollment classes. While our data are not as rich as interviews, they are substantially more complete than pretesting and post-testing will allow (in the limits of very large enrollments).

In this paper, we present two variations on the RCM, contrast them to prior work, and present data on the dynamism of student understanding of Newton's third law (N3L).

Newton's third law is an especially interesting topic to study throughout the introductory physics curriculum. A typical mechanics class, the first of a physics sequence, has a fairly canonized topic order: kinematics, forces, energy, momentum, torque, and angular quantities. N3L is thus introduced in the second unit of the course, allowing for a few weeks of preinstruction data collection. After the forces unit, a seemingly unrelated topic allows students' ideas about N3L to relax back towards the preinstruction state. In the momentum unit, N3L is again emphasized (usually in the context of collisions). There are typically a few weeks remaining in the term after the momentum unit, allowing for postinstruction data collection in addition to the two opportunities during instruction. Past research on students' understanding of N3L has focused on the differences between different instructional methods [10–13] or N3L as a convenient topic for investigating deeper issues in student understanding [8,14–16]. This paper follows the latter tradition. A first course on electromagnetism typically follows the mechanics course. Even though N3L is still important in electromagnetic contexts, the law is emphasized less during instruction (which commonly

Published by the American Physical Society under the terms of the Creative Commons Attribution 3.0 License. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

focuses on fields instead of interaction forces). Tracking students' understanding of N3L during an electricity and magnetism (E&M) course should reveal longer-term effects of instruction as well as any interference effects between E&M material and older N3L ideas.

II. THEORETICAL FRAMEWORK

The field of cognitive psychology offers many reasons to suspect that student performance is complicated and time dependent. Learning curves, such as those predicted by the Rescorla-Wagner model [3], show that for repeated training, scores increase quickly at first, then level off. This simple error-reduction model assumes that the amount of learning depends on the difference between perfect and actual performance, and matches empirical data in a wide array of simple learning tasks in a variety of species, including humans. ''Forgetting curves,'' such as those studied by Ebbinghaus [17], reveal that memory performance decays exponentially after training ceases, eventually reaching a new minimum value, and have been demonstrated in a wide variety of tasks and time scales (from seconds to decades).

A third phenomenon, interference, occurs when two pieces of related information (or tasks) are learned. Performance on one can significantly decrease when the second is learned either before (proactive) or after (retroactive), and the amount of interference increases with the degree of similarity between the two pieces of information [1,4]. For example, student performance on questions involving the vector superposition of electric fields falls below preinstruction levels during subsequent instruction on the scalar concepts of electric potential and circuits [5]. Responses can be brought back up to their peak levels during instruction on the vector-based topic of magnetic fields. However, electric and magnetic fields have different effects on charges—though they are both vector fields—and a similar degradation can be seen in students' understanding of electric forces during the magnetic fields unit [18].

Students' understanding of these related, but different, phenomena is decidedly nonmonotonic. There is good evidence to suggest that nonmonoticity of performance and understanding is a universal human learning phenomena which is not limited to the university physics classroom. (For excellent overview articles, see Siegler [19].) The common pretesting and post-testing is woefully insufficient to investigate the dynamism of student understanding. Furthermore, given the short time scales over which scores may change dramatically, changing the timing of either pretests or post-tests may artificially alter calculated gains.

Instructors are not blind to these effects: conventional wisdom states that students will forget much after an exam. From another perspective, many students believe that cramming before an exam will increase their score on the exam. Using the RCM, we can test whether instructor's pedagogical content knowledge on this topic is supported by data, a heretofore unstudied proposition.

Phenomenologically, we divide postinstruction changes in performance into two categories: that which is characterized entirely by negative slope (though often positive second derivative, as with exponential death) and that which is characterized by changing signs of slopes, i.e., bumps or dips. In this paper, we focus on changes of the second sort, noting that their effects may be dwarfed by changes of the first sort. Because these bumps or dips always occur concurrently with instruction on related but different topics, we conceptualize them as an interference effect between older and newer material [20].

This phenomenological description may have several underlying causes. As a trivial example, consider that performance on these tasks may be the combined result of several independent skills which increase or decrease monotonically at different rates and intercepts. Superposition could produce an apparent series of bumps and dips. While it is mathematically possible to model our data in this manner, the number of free parameters is quite large and we seek a more parsimonious model.

It is also possible that new ideas may be improperly extended to cover old material, causing destructive interference between old and new. The interference may resolve itself when the new material is properly limited in scope or forgotten altogether. In these cases, we can expect that performance afterwards returns to nearly the same level as beforehand. Alternately, new ideas may simply be difficult to incorporate and thus temporarily ''crowd out'' older ones as part of the normal process of conceptual change [21]. During the change, a temporary dip is followed by a more durable bump in both old and new material. The former explanation reduces to low connectedness [22] learning, while the latter reduces to high connectedness. Both explanations fit learning under the resources model via plasticity [23], while the latter is more amenable to horizontal and vertical conceptual reorganization [24]. Without access to ongoing data over many weeks or years (and perhaps in-depth qualitative data as well), it is not possible to distinguish between these explanations. We limit ourselves to a phenomenological approach here.

III. METHODS

A. Population

This study took place at a northeastern, large, four-year private university with high undergraduate enrollment and no graduate program in physics, code named the ''Institute.'' The year is divided into four 10-week quarters (including a summer term).

Each year $\approx$ 2400 students take introductory calculus-based physics, which is offered in a workshop format that integrates lecture, experiment, and short group activities. Adapted after the SCALE-UP project [25], the classes meet for three 2-hour sessions each week, with students seated at

TABLE I. Study population for fall and winter divided by course. In the fall, the primary course is E&M, and in the winter, the primary course is Mechanics.

	Fall		Winter		Total
Course	No. of section	N	No. of section	N	N
Mechanics	5	142	14	441	583
E&M	8	257	5	144	401

TABLE II. Demographic data for students taking the tasks. Mechanics is typically taken in the winter quarter of students' first year. Students typically take E&M in the fall of their second year. Women make up 14%–24% of the students.

	Gender	Year		
	(Male/female)	1	2	3–6
Fall 2009 Mechanics	117/20	26	81	47
Winter 2009 Mechanics	337/89	393	21	24
Fall 2009 E&M	201/51	1	201	54
Winter 2009 E&M	107/34	1	110	39

tables of six and working in small groups. Classrooms accommodate up to 42 students, with enrollment in each section varying. Engineering students dominate the population, comprising 57%–83% of the students in Mechanics and 65%–83% of E&M students. Most students begin the sequence in the winter of their freshman year.

Our study has two phases. In the first phase, we tested students in the first (Mechanics) and third (E&M) courses during the fall and winter quarters of the 2009-2010 academic year. In the second phase, we tested students in all three courses in all three quarters of the 2010-2011 school year. The syllabus is unchanged, so we collapse our data across multiple quarters (by week) to increase sample size. Participation for each course and phase is summarized in Table I.

The majority of the students begin the 3-quarter sequence in the winter of their freshman year, concluding with E&M in the fall of their second year. Second-year and older students, taking the course later in their career than normal, make up 81% (128/139) of the off-sequence Mechanics class (see Table II).

B. Methodology

We use variants of the RCM developed by Sayre and Heckler [5]. The RCM allows researchers to probe student understanding via short quizzes administered frequently throughout a course or series of courses, subject to the following constraints:

(a) Each conceptual topic is tested in only one task, to avoid false isomorphism between different questions which purport to be about the same topic.
(b) Each student is tested on each task only once, to avoid test-retest effects.
(c) Each topic is tested every time period (week, day, month, etc.).

A "task matrix" which satisfies these constraints is constructed at the beginning of each term. To explore them, consider an m-week course in which n students enroll and we wish to investigate student understanding on a weekly basis. If we would like each student to take a task every week (so that it becomes habit), then by constraint 2 we need ℓ different tasks, where $\ell \geq m$. To test each task every week (constraint 3), then we need at least ℓ groups of students, each of which has N/ℓ members. In the limit of large enrollments, the random assignment of students to groups means that (statistically) the groups are indistinguishable from each other, and the only differences between groups in performance on a specific task is due to the effects of instruction (because groups are tested in different week per constraint 2). Creating a task matrix can be a nontrivial task; however, it is not insurmountable.

Of course, more frequent testing is possible: in the original RCM, testing occurred 3–5 times each week [6], with a concomittant loss of statistical power (given the same total population). Less frequent testing is also possible if enrollment numbers do not support weekly testing.

The original RCM required that every student visit the research laboratory for 1 hour of testing. This is extremely time intensive for researchers, who must sign up, remind, proctor, and (in some cases) reschedule hundreds of students in a quarter. After proctoring, researchers must digitize the paper-based data in preparation for analysis. In this study, we tested two variants to the RCM which are intended to reduce the overhead to testing.

TABLE III. Methodology affordances and limitations.

Feature	Original RCM	Phase 1	Phase 2
Each student participates	Once for 1 hour	Every week for 5-10 minutes	
Instructors	Mention the project in class	Administer quizzes weekly	Mention the project in class
Groups are determined	By which students participate each week	By enrollment in different course sections	Randomly at the start of term
Data are collected	On paper in the research lab	On paper during class	Online, unproctored
Data are digitized and reduced	By hand, laboriously		Automatically
Reports of student progress are available	After the term		Instantly

In both RCM variants, each student participates every week for 5–10 minutes instead of once for 1 hour. The details of these changes are summarized in Table III.

1. Phase 1: Paper

The Institute's system of many sections per course makes it ideal for a paper-based between-student study. In the first phase of our study, we administered the tasks on paper each week in class. Because students are already in class, researchers do not have to schedule and proctor them in special sessions, and students do not have to remember to attend a special research session.

The groups have similar grade, cumulative grade point average (GPA), major, and gender distributions. (The study design precludes testing multiple groups on N3L at the same time.) Students had $\approx$ 10 minutes to complete each task, which were sometimes appended to an instructor-generated quiz.

This method still requires significant researcher time in data collection and digitization for researchers. It requires a small amount of class time every week, which some instructors might be loathe to give up. Furthermore, because each course section is a different group, it is possible that some instructors are better than others, artificially skewing the data in some weeks. While the phase 1 design cannot test the instructor effect easily, the phase 2 design (detailed below) can test it easily, and has found it to be not statistically significant.

2. Phase 2: Electronic

To enhance the method's applicability to larger classes and substantially streamline early analysis, we developed a web-based testing system. The rapid analysis and web reports (RAWR) system automates task administration and is accessible from any modern web browser. It further simplifies subject pool management and data collection and reduction for researchers. For instructors, it reduces the strain on class time as well as providing real-time reports of their classes' learning and forgetting which can be used to modify instruction.

The RAWR system runs on an entirely open-source and free platform. The student and instructor interfaces are web based and interface with the Institute's online course delivery system (myCourses). Students see it as an extension of a familiar Institute system, using the same user name and password that they do for all Institute-related systems. Demographic information, informed consent, and task results are stored in a relational (mySQL) database for easy retrieval. Reports, statistics, and graphs are generated using php scripts that interface with R, an open-source statistics package. This allows us to automate many different types of analyses, including questions useful to researchers as well as instructors.

Instructors announce the RAWR project to their students early in the quarter. Instructor policy varies, with some giving small amounts of participation credit and heavily stressing their commitment to the project. Others mention it only a few times, and give no credit. Confirming that participation is independent of instructor is therefore an important first step to validating the system, and is described below.

Students register on our site and are first asked for informed consent. Their participation is still expected (by their instructor) regardless, but data from students who do not consent are not included in this study. Approximately 8% of students either withhold consent or drop the course after the second week and are not included. Students also provide demographic data such as gender, major, and prior math and physics courses, which are correlated with later test results.

Students are assigned into 10 groups per course. Group membership is randomly assigned, with each group having 25–45 students. Each group takes a different task every week, and each task is assigned to a different group every week. Students cannot see which task they will take until the time it opens for completion. Typically, tests become available on Monday of each week and are due on Wednesday. Students may log in at any time to complete the task, although there is a 30 minute time limit once they start the test. We record both their starting time and completion time to test whether students take enough time to read the questions entirely. RAWR automatically Emails students each week to remind them to take the tasks as they become available.

When students complete tasks on RAWR, it records their responses, the time that they accessed the task, and the time of completion. While students are permitted to take up to 30 minutes to complete a task, the tasks are designed to be completed in 5–10 minutes. Any response which takes less than 30 seconds to complete is removed from further analysis.

Because groups are statistically independent, we can compare the performance of different groups across weeks, essentially capturing student understanding on a weekly time scale. A time plot of average performance, termed the *response curve*, is sensitive to the particulars of the week—the current topic of instruction and coincidence with exams or homework. The conventional pretest and post-test corresponds to the first and last points on the curve, and can miss much of the dynamic evolution of understanding.

Participation in phase 2 is significantly lower than in phase 1. Roughly half the students take seven or more tasks in phase 2. When the tasks were given as in-class quizzes, participation each week was closer to 85%, but we did not track individual student participation.

IV. DEMOGRAPHICS AND PARTICIPATION

It is possible that students' demographics affect their score on N3L; it is also possible that selection effects of

which students participate in the study could skew our results.

We collect demographic information from students about gender, year in school, major, and prior physics courses. Additional information is available from the Registrar of the Institute, including cumulative GPA and course grade (on completion of the course).

A. Participation rates

In phase 1, approximately 95% of enrolled students participated in the study each week. With participation rates that high, no further analysis of selection effects are warranted. In phase 2, participation rates dropped substantially. Analysis of variance (ANOVA) tests were conducted to determine whether the course (Mechanics, Waves, or E&M), section, professor, or major were significantly correlated with the participation rate in phase 2. For the fall and winter quarters, none of these factors was significant, with $p > 0.1$ for all combinations. Course and professor were marginally significant ($p = 0.035$) for the spring quarter, although the professors for this quarter were, for the most part, the same as previous quarters.

The grade achieved in the course was statistically significant ($p < 0.001$), as was the student cumulative GPA at the time of taking the course. This latter can be seen in Fig. 1, which groups students within a particular rate by GPA. Note the significantly larger percentage of students with a GPA less than 2.5 (bottom two segments in each bar) who answer 0, 1, or 2 tasks. This correlation is also shown in Fig. 1, where GPA is plotted versus the number of completed tasks.

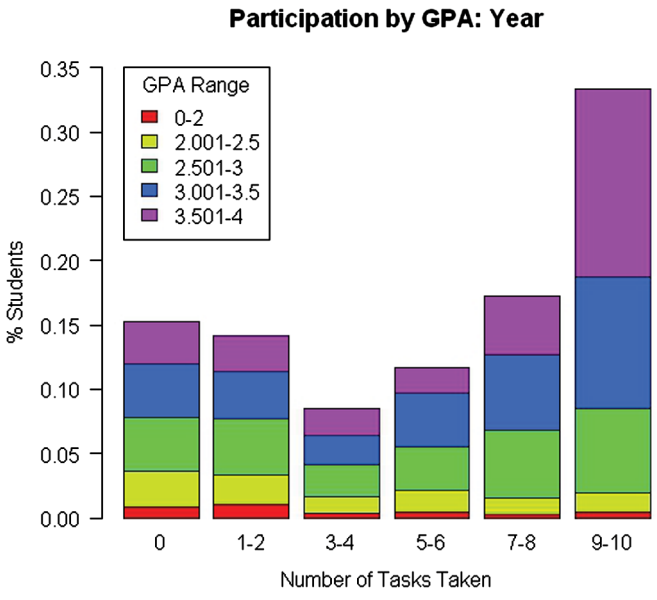


FIG. 1 (color online). Student participation is bimodal, with 40% of students completing seven or more tasks, and ~30% of students completing less than three.

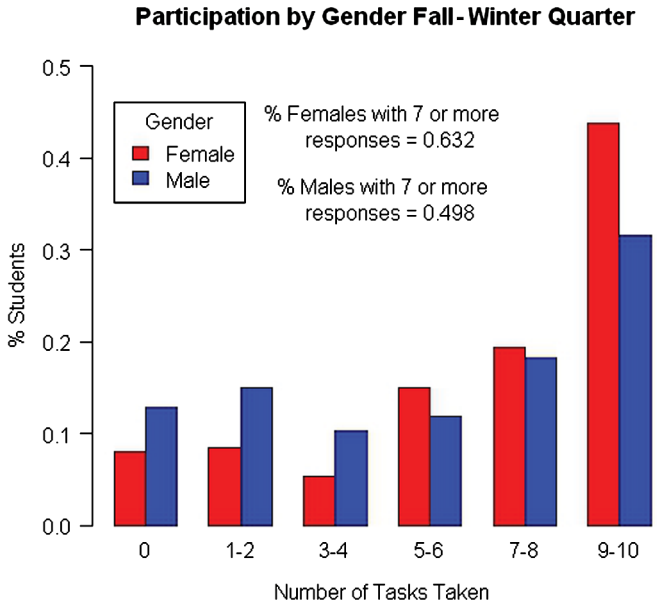


FIG. 2 (color online). Generally, women participate more often than men. The course is about 70% male.

We also notice a small gender bias, with women slightly more likely to participate, as shown in Fig. 2. Sixty-three percent of women took 7 or more tasks, compared with only 50% of men. This was statistically significant in the fall and winter quarters (as shown); the spring quarter showed a smaller gap (women, 58%; men, 50%) that was not statistically significant ($p = 0.1$).

B. Gender, major, and GPA

To measure demographic effects, we use multilevel modeling to estimate the influence of gender, major, and GPA on students' scores. Additionally, we test whether students' high school physics courses have an effect on their N3L knowledge in Mechanics.

We are interested in how majors and genders may have different impacts on the students' performances on the task. Therefore, we need to include both student-level and group-level indicators, but classical regression either tends to ignore group effects or it may include group effects while still ignoring interactions among groups. However, the demographic group sizes in our study vary a lot. For example, we have about 100 physics majors but only several psychology majors. Simply using the local information is fraught if the sample size is small in the group, but regression ignoring group indicators can be misleading in ignoring group-level variation. Multilevel modeling allows the estimation of group averages and group-level effects, compromising between the overly noisy within-group estimate and the oversimplified regression estimate that ignores group indicators [26].

To better estimate the effect of demographic groups with small populations (for example, psychology majors), we

use a Bayesian approach with Gibbs sampling that averages over the uncertainty in all the parameters of the model [27]. We choose to work with simulations (rather than simply point estimates of parameters) because we can directly capture inferential uncertainty and propagate it into predictions. Whenever we represent inferences for a parameter using a point estimate and standard error, we are performing a data reduction. If the estimate is normally distributed, this summary discards no information because the normal distribution is completely defined by its mean and variance. But in other cases it can be useful to represent the uncertainty in the parameter estimation by a set of random simulations that represent possible values of the parameter vector (with more likely values being more likely to appear in the simulation). By simulation, then, we mean summarizing inferences by random numbers rather than by point estimates and standard errors [26].

We divide the students' GPA into five groups. Gender is coded as a binary. In phase 1, students' majors are collapsed into five categories: computer science, engineering, physics, science, and other. In phase 2, the RAWR system permitted the collection of more detailed information about majors, and major information is left uncollapsed across the 55 majors who take introductory physics at the Institute.

We use a linear regression,

$$y_i \sim N(\mu + \gamma_{\mathrm{GPA}[i]} + \delta_{\mathrm{major}[i]} + \beta x_i, \sigma_y^2), \qquad (1)$$

where y_i represents the score of the ith student, μ is the average expected score on the instrument, $\gamma_{\mathrm{GPA}[i]}$ and $\delta_{\mathrm{major}[i]}$ are group-level indicators specifying different GPA groups and majors, β is a constant, and x_i is an individual-level indicator of gender. Thus, if a male ($x_i = 1$) student with a GPA of 3.561 ($\gamma_i = A$) who is a physics major ($\delta_{\mathrm{physics}}$) takes the test, his score would be

$$y = \mu + \gamma_A + \delta_{\mathrm{physics}} + \beta \cdot 1. \qquad (2)$$

We find that the average expected scores (μ) are 3.0 ± 0.5 in phase 1 and 3.5 ± 0.5 in phase 2. That phase 2 has a larger expected score could be because of selection effects in phase 2, where we oversampled better students, or it could stem from the format shift from paper-based to computer-based testing. This is discussed in more detail in Sec. IV A.

In both phases, we find a slight effect of cumulative GPA. Students who score better have higher GPAs (Table IV). We find a substantial gender effect. Men score better by 1.9 ± 0.3 questions in phase 1 and by 0.9 ± 0.2 questions in phase 2. In phase 1, we do not find an effect of major (collapsed into five options). In phase 2, where the data specify majors in more detail, we find that most majors (of the 55 enrolled in physics) score about the same as each other. Only five majors score significantly better or worse than the average across all majors (Table V).

TABLE IV. Effects of GPA on N3L score.

GPA	Phase 1 mean ± SD	Phase 2 mean ± SD
0–2	-0.2 ± 0.6	-1.3 ± 0.6
2–2.5	0 ± 0.4	0 ± 0.5
2.5–3	-0.2 ± 0.4	-0.2 ± 0.5
3–3.5	-0.1 ± 0.4	0.2 ± 0.5
3.5–4	0.5 ± 0.4	0.8 ± 0.5

TABLE V. Effect of major in phase 2. For brevity, we list only the majors with nonzero coefficients.

	mean ± SD
Computer Science	-0.3 ± 0.2
Aerospace Engineering	0.7 ± 0.3
Physics	0.8 ± 0.3
Mechanical Engineering Technology	-0.5 ± 0.4
Biomedical Engineering	-0.4 ± 0.3

C. Lasting effects of prior physics instruction

We divide the students who are currently enrolled in Mechanics into four different groups based on their answers to the question "What is the highest level physics course for which you have credit?"

As Mechanics is the first course in the sequence, the vast majority of students respond about the physics classes they took in high school. (A trivial minority never took physics before, or took a different university course, such as algebra-based physics; both of those populations are excluded from the analysis in this section.) The four possible high school physics classes are "regular physics," advanced placement (AP) physics B, AP physics C, and international baccalaureate (IB) physics. Data are self-reported, and we neither ask students for their scores on the AP or IB exams (where applicable) nor assess the quality or content of their prior physics classes.

We use a linear model where

$$y_i \sim n(\alpha_{\mathrm{group}[i]} + \beta x_i, \sigma_y^2), \qquad (3)$$

where y_i, β, and x_i have the same meanings as in the previous question, and group$[i]$ is one of the four previous physics classes. Here we have collapsed across major and GPA to look primarily at the effects of gender and prior

TABLE VI. Effects of prior physics classes.

Course	Phase 1 mean ± SD	Phase 2 mean ± SD
Regular physics	2.8 ± 0.4	3.3 ± 0.5
Physics AP B	2.6 ± 0.5	3.5 ± 0.4
Physics AP C	4 ± 0.7	3.5 ± 0.5
IB physics	2 ± 1.2	3.4 ± 0.4

physics preparation. Again, we use WINBUGS to run 3 Markov chains in parallel with 4000 iterations. Data are summarized in Table VI.

Phase 1 data suggest that there is little, if any, effect of prior physics classes on students' understanding of N3L. It is possible that the selection effects in phase 2 have washed out any effect of prior physics classes in this analysis.

V. PHASE 1: NEWTON'S THIRD LAW TASKS

Multiple-choice tasks were devised to probe student understanding of Newton's third law. In phase 1, two tasks of four questions each were developed. In order to align with the instruction of the different classes, tasks were couched in appropriately different contexts. Sample tasks are shown in Figs. 3 and 4.

In Mechanics, the task involved a car pulling a trailer (see Fig. 3). Students are asked to compare the forces acting on the car and trailer as the car speeds up, travels at constant speed up a hill, travels at constant speed on a level road, and slows down. Students choose one of the given answers (A–E). The answer choices for each question were the same and the students could select each answer as many times as they wanted. These questions were chosen to cover the space of commonly occurring student models in pulling scenarios [10].

Which of the below answers (A-E) indicates the relation between the car and trailer in the figure if:

1) the car is initially stopped and begins to move.
2) The car is driven at constant speed up a hill.
3) The car is slowing down.

A: The trailer pulls on the car a lot, but the car doesn't pull on the trailer.

B: The trailer pulls on the car more than the car pulls back on the trailer, but the car still pulls on the trailer.

C: The trailer pulls on the car exactly as much as the car pulls on the trailer.

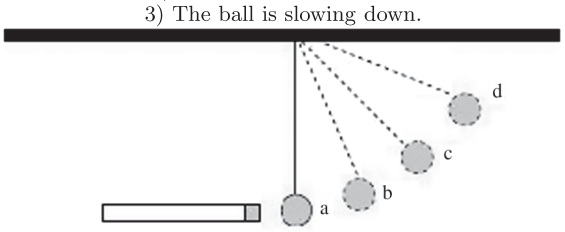
D: The car pulls on the trailer more than the trailer pulls on the cal, but the trailer still pulls on the car.

E: The car pulls on the trailer a lot, but the trailer doesn't pull on the car.

FIG. 3. Phase 1 prompt and responses for N3L task for Mechanics. The students were asked to consider situations where the car was speeding up, constant speed up a hill, constant speed on a level road, and slowing down. In each case, the force that the car exerts on the trailer is equal and opposite to the force that the trailer exerts on the car.

Which of the below answers (A-E) indicates the relation between the rod and ball in the figure if:

1) the ball is initially stopped and begins to move.
2) The ball is speeding up.
3) The ball is slowing down.



A: The rod pushes on the ball a lot, but the ball doesn't push on the rod.

B: The rod pushes on the ball more than the ball pushes back on the rod, but the ball still pushes on the rod.

C: The rod pushes on the ball exactly as much as the ball pushes on the rod.

D: The ball pushes on the rod more than the rod pushes on the ball, but the rod still pushes on the ball.

E: The ball pushes on the rod a lot, but the rod doesn't push on the ball.

FIG. 4. Phase 1 prompt and response for N3L task for E&M. The students were asked to compare the forces on the rod and ball for the three scenarios where the ball started to move, sped up, slowed down, and at rest at the apex of the swing. As in the mechanics example, the forces in each case are equal and opposite.

For E&M, the task was rewritten to involve electric charges (see Fig. 4). Students compared the forces acting on the rod and ball as the ball starts to move, speeds up, and slows down as it swings away from the rod and finally when it comes to rest. This question is not completely isomorphic to the Mechanics formulation, and so we do not directly compare the Mechanics and E&M responses. Rather, we look for changes in the response over the course of each quarter, and similarities in how this behavior corresponds to the topic of instruction.

Between-student data are collected by having different groups of students take the tasks (e.g., Figs. 3 or 4) each week. Groups corresponded to different sections of the same course; group sizes ranged from 13 to 42 students. We group all incorrect student answers together and plot the percentage of students getting the correct response in Figs. 5 and 6. Error bars in Figs. 5 and 6 are 1σ, calculated from a binomial distribution.

VI. RESULTS

A. Instruction's positive impact

Figure 5 shows the response curve for students in the Mechanics course. Shown are average responses for the three nontrivial questions asked in Fig. 3; a question involving the car traveling at constant speed shows a ceiling effect where almost all students answer correctly independent of week, instructor, or any other variable. Although

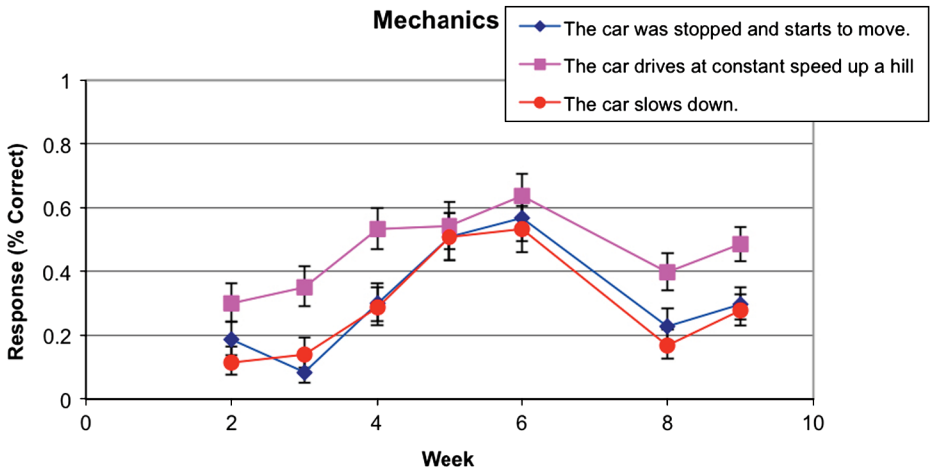


FIG. 5 (color online). Response curve for Mechanics. Before instruction response could be chance. There is a broad peak during instruction with a maximum during week 6 which is the end of the section on forces.

80% of students have taken physics prior to the introductory course at the Institute, response during the first few weeks of the course, before explicit instruction of forces or Newton's laws, hovers around the chance line of 20%. Instruction on forces begins in week 4, and student performance begins to rise, culminating with a maximum performance in week 6. Week 6 is also the last week of instruction on forces, and includes the examination. After instruction, the response rapidly drops, with two of the questions ending just above the chance line at the end of the quarter.

B. Instruction's negative impact

Figure 6 shows the response curve for all students in the E&M course, a question involving the ball at rest having been omitted due to the presence of a ceiling effect. At the Institute, E&M is typically taken in the fall quarter. This means that because of the summer break, it has been approximately 5 months since these students last saw instruction on forces and Newton's laws. (The second quarter deals with rotational motion, waves, and miscellaneous physics topics.) Nevertheless, students enter with an initial response of 66%, significantly higher than they exited Mechanics. We have three potential explanations. Most likely is a winnowing effect, with the weakest students leaving the sequence before reaching E&M. Failure rates (defined as obtaining a D, F, or withdrawing) in Mechanics average around 25%, and an additional $\approx 17\%$ exit between Mechanics and E&M. Therefore, students entering E&M are the top 62% of the students in Mechanics, and a higher performance is expected. Less likely is the possibility that instruction in the second quarter has bolstered student understanding. It is also possible that the two scenarios do not appear similar to students.

The most significant feature of Fig. 6 is the pronounced dip in week 4 to 41%. This drop, 25% points below the average, cannot be explained by instructor or section variance, and so we assert that course topic is the most likely

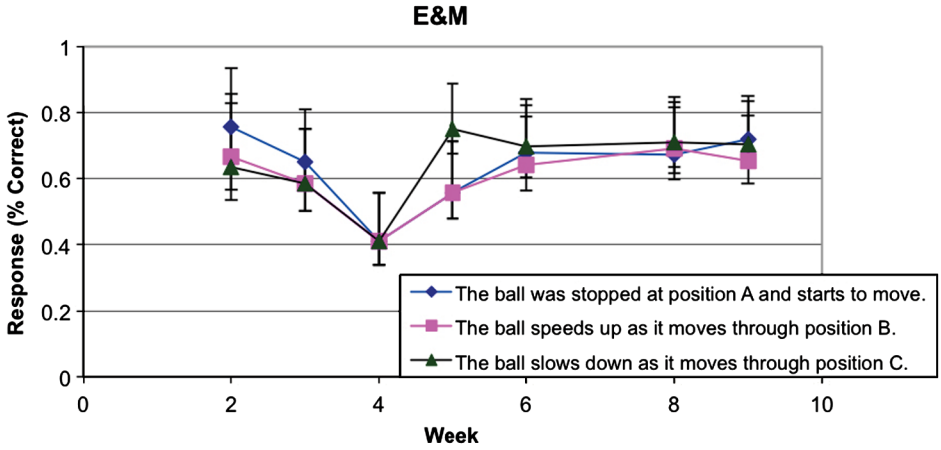


FIG. 6 (color online). Response curve for E&M. The response is mostly flat around the average of 66% with a measurable dip during week 4. This dip corresponds to the period of instruction in electric potential.

cause. In E&M, the first three weeks are spent on electric fields, Coulomb's law, and Gauss's law. Week 4 shifts the topic from vector-based concepts to the scalar topics of electric potential and voltage. We speculate, therefore, that instruction on the scalar electric concepts interferes with response to a vector-based (Coulombic force) question. In week 5 instruction shifts to current, resistance, and circuits. While this is also scalar based, and we note that the week 5 performance is still below average, we suspect that because instruction is not explicitly involving electric charges the interference effect is lessened. These topics are amenable to further study, and the richness of data in interviews would be appropriate here.

VII. SIGNIFICANCE

It has been established [5] that student understanding is dynamic and time dependent. In this study we have shown that this dynamism continues far beyond the immediate period surrounding instruction. Inasmuch as long-term studies of student understanding exist (for examples, see the study of E&M among juniors by Pollock and Chasteen [28] or the study of non-STEM (Science, Technology, Engineering, and Mathematics) majors upon graduation by Barrantes *et al.* [29]), they support the long-term behavior of our data. While long-term studies are important, they obscure the dynamism present on the scale of weeks.

Student response to questions on vector-based topics, like Newton's third law, are sensitive to *any* physics instruction they are receiving at the time. ''Dissonant'' instruction, e.g., topics that emphasize a scalar concept, suppresses student scores. It is fortunate that this interference disappears when instruction returns to more ''consistent,'' i.e., vector-based, topics.

The impact of current instruction on previously learned knowledge has been loosely termed ''interference'' [4]. It underscores the complexity of student learning, as students struggle to identify, activate, and use appropriate knowledge in response to a prompt. Even strong students, who have already progressed through two previous quarters of physics and show a high initial score, struggle to reconcile a strange prompt with their current frame of mind. The implications for testing and assessment may be profound, calling into question the accuracy of any single evaluation. Phenomenologically, we note the existence of the interference; interview or in-class video-based observational data of students may help pinpoint more specific causes. Subsequent research will look at interference effects in strong and weak students, mainstream, and remedial sections, and in more explicit vector tasks.

ACKNOWLEDGMENTS

We thank Gordon Aubrecht and Andrew Heckler for their assistance in task development; Conor Frame assisted with data analysis; Sam Milton, Tyler Koch, and Nathan Popham developed the RAWR system for phase 2. This work was supported by the Rochester Institute of Technology Physics Department Capstone Research program, and we acknowledge the department for travel support. This work is partially supported by NSF DUE Grants No. 0941889 and No. 0941378, and by the Committee for Undergraduate Research at Wabash College.

[1] L. Postman and B. J. Underwood, Critical issues in interference theory, Mem. Cogn. **1**, 19 (1973).

[2] G. B. Semb, J. A. Ellis, and J. Araujo, Long-term memory for knowledge learned in school, J. Educ. Psychol. **85**, 305 (1993).

[3] R. A. Rescorla and A. R. Wagner, A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement,in *Classical Conditioning II: Current Theory and Research*, edited by A. H. Black and W. F. Prokosy (Appleton-Century-Crofts, New York, 1972), pp. 64–99.

[4] M. E. Bouton, Context, time, and memory retrieval in the interference paradigms of Pavlovian learning, Psychol. Bull. **114**, 80 (1993).

[5] E. C. Sayre and A. F. Heckler, Peaks and decays of student knowledge in an introductory E&M course, Phys. Rev. ST Phys. Educ. Res. **5**, 013101 (2009).

[6] A. F. Heckler and E. C. Sayre, What happens between pre- and post-tests: Multiple measurements of student understanding during an introductory physics course, Am. J. Phys. **78**, 768 (2010).

[7] M. A. Kohlmyre, M. D. Cabellero, R. Catrambone, R. W. Chabay, L. Ding, M. P. Haugan, M. J. Marr, B. A. Sherwood, and M. F. Schatz, Tale of two curricula: The performance of 2000 students in introductory electromagnetism, Phys. Rev. ST Phys. Educ. Res. **5**, 020105 (2009).

[8] R. K. Thornton and D. R. Sokoloff, Assessing student learning of Newton's laws: The Force and Motion Conceptual Evaluation and the Evaluation of Active Learning Laboratory and Lecture Curricula , Am. J. Phys. **66**, 338 (1998).

[9] 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 (1998).

[10] T. I. Smith and M. C. Wittmann, Comparing three methods for teaching Newton's third law, Phys. Rev. ST Phys. Educ. Res. **3**, 020105 (2007).

[11] M. Sahin, Effects of problem-based learning on university students' epistemological beliefs about physics and physics learning and conceptual understanding of Newtonian mechanics, J. Sci. Educ. Technol. **19**, 266 (2010).

[12] M. Kocakülah, Development and application of a rubric for evaluating students' performance on Newton's laws of motion , J. Sci. Educ. Technol. **19**, 146 (2010).

[13] D. Fraser and C. Linder, Teaching in higher education through the use of variation: Examples from distillation, physics and process dynamics, Eur. J. Eng. Educ. **34**, 369 (2009).

[14] L. Bao, K. Hogg, and D. Zollman, Model analysis of fine structures of student models: An example with Newton's third law, Am. J. Phys. **70**, 766 (2002).

[15] R. E. Scherr and E. F. Redish, Newton's zeroth law: Learning from listening to our students, Phys. Teach. **43**, 41 (2005).

[16] S. Ramlo, Validity and reliability of the force and motion conceptual evaluation, Am. J. Phys. **76**, 882 (2008).

[17] H. Ebbinghaus, *Memory: A Contribution to Experimental Psychology* (Columbia University, New York, 1913).

[18] T. Scaife and A. Heckler, Interference between electric and magnetic concepts in introductory physics, Phys. Rev. ST Phys. Educ. Res. **7**, 010104 (2011).

[19] R. Siegler, U-shaped interest in U-shaped development—and what it means, J. Cognit. Dev. **5**, 1 (2004).

[20] We have expanded slightly the cognitive science concept of ''interference'' (which is purely destructive) to align better with the physical concept of ''interference'' (such as of waves) which can be constructive or destructive. Our modifications fit a broader range of data and can be explained using similar causal mechanisms.

[21] R. E. Scherr and M. C. Wittmann, The challenge of listening: The effect of researcher agenda on data collection and interpretation, in *Proceedings of the Physics Education Research Conference, Boise, ID, 2002*, edited by S. Franklin, K. Cummings, and J. Marx, http://www.compadre.org/per/items/detail.cfm?ID=4298.

[22] D. Pritchard, Y.-J. Lee, and L. Bao, Mathematical learning models that depend on prior knowledge and instructional strategies, Phys. Rev. ST Phys. Educ. Res. **4**, 010109 (2008).

[23] E. C. Sayre and M. C. Wittmann, Plasticity of intermediate mechanics students' coordinate system choice, Phys. Rev. ST Phys. Educ. Res. **4**, 020105 (2008).

[24] D. McBride, D. Zollman, and N. Rebello, Method for analyzing students' utilization of prior physics learning in new contexts, Phys. Rev. ST Phys. Educ. Res. **6**, 020101 (2010).

[25] R. J. Beichner, J. M. Saul, R. J. Allain, D. L. Deardorff, and D. S. Abbott, Introduction to scale-up: Student-centered activities for large enrollment university physics, in *Proceedings of the Annual Meeting of the American Society for Engineering Education, Seattle, Washington, 2000*, http://www.ncsu.edu/PER/Articles/01ASEE_paper_S-UP.pdf.

[26] A. Gelman and J. Hill, *Data Analysis Using Regression and Multilevel/Hierarchical Models* (Cambridge University Press, Cambridge, England, 2006).

[27] Mechanically, we use the R software package together with WINBUGS, which runs 3 Markov chains with 4000 iterations each.

[28] S. J. Pollock and S. V. Chasteen, Longer term impacts of transformed courses on student conceptual understanding of E&M, AIP Conf. Proc. **1179**, 237 (2009).

[29] A. Barrantes, A. Pawl, and D. E. Pritchard, What do seniors remember from freshman physics?, AIP Conf. Proc. **1179**, 47 (2009).