What do seniors remember from freshman physics?

Andrew Pawl,* Analia Barrantes, and David E. Pritchard
Physics Department, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

Rudolph Mitchell
Teaching and Learning Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA
(Received 24 August 2012; published 10 December 2012)

We have given a group of 56 Massachusetts Institute of Technology (MIT) seniors who took mechanics as freshmen a written test similar to the final exam they took in their freshman course as well as the Mechanics Baseline Test (MBT) and the Colorado Learning Attitudes about Science Survey (CLASS). Students in majors unrelated to physics scored 60% lower on the written analytic part of the final than they would have as freshmen. The mean score of all participants on the MBT was insignificantly changed from their average on the posttest they took as freshmen. However, the students' performance on 9 of the 26 MBT items (with 6 of the 9 involving graphical kinematics) represents a gain over their freshman posttest score (a normalized gain of about 70%), while their performance on the remaining 17 questions is best characterized as a loss of approximately 50% of the material *learned* in the freshman course. On multiple-choice questions covering advanced physics concepts, the mean score of the participants was about 50% lower than the average performance of freshmen. Although attitudinal survey results indicate that almost half the seniors feel the specific mechanics course content is unlikely to be useful to them, a significant majority (75%–85%) feel that physics does teach valuable problem solving skills, and an overwhelming majority believe that mechanics should remain a required course at MIT.

DOI: 10.1103/PhysRevSTPER.8.020118 PACS numbers: 01.40.Di, 01.40.Fk

I. INTRODUCTION

A. Goals

We have studied the physics knowledge of graduating seniors who took introductory Newtonian mechanics during their freshman year. Motivations for this study included determining what knowledge is retained (or improved) and whether conceptual knowledge is retained better than analytic knowledge. We also wanted to investigate what aspects of the students' subsequent behavior (e.g., their academic major, participation in tutoring for freshman physics, etc.) influenced retention. Finally, it was an opportunity to look for evolution in the students' attitudes toward learning physics.

B. Sample and procedure

Our sample consisted of students who took and passed (grade of C or better) the regular freshman mechanics course (MIT 8.01) in the fall of 2005, and who were still enrolled at Massachusetts Institute of Technology (MIT) in spring 2009. Students were recruited by Email and informed that they would be retaking a final exam from a standard freshman course, but not told that the subject was physics. (An informal poll of the participants indicated that a majority had guessed the subject was physics, but none had studied for the retest based upon this suspicion.) Students were guaranteed \$75.00 for spending at least 3 hours on the materials and offered a 1/3 chance of receiving a performance-based award of an additional \$100.00. A total of 56 students out of 486 invited participated in the retest. The breakdown of the participants by freshman course grade was a good approximation to the distribution for all 506 students who took the mechanics course in fall 2005 ($p = 0.94$ for equivalent distributions when binned by letter grade).

The students were allowed up to time and a half (4.5 h total) to complete the test materials, which consisted of

- An 18-question participant survey ($\sim$15 min).
- The Colorado Learning Attitudes about Science Survey (CLASS) [1] standard survey ($\sim$10 min).
- The Mechanics Baseline Test (MBT) [2] standardized mechanics test ($\sim$45 min).
- A final exam composed of 7 multiple-choice and 4 written problems ($\sim$2–3 h).

II. METHODS

A. Division of the sample by major

One goal of this study was to compare the retention of the mechanics curriculum by students in various majors. To gain statistical leverage, we classified the majors into three

*Now at: Department of Chemistry and Engineering Physics, University of Wisconsin-Platteville, Platteville, Wisconsin, USA.
pawla@uwplatt.edu

groups. Group 1 encompassed majors that were least related to physical science, and therefore least likely to use or review the content of freshman mechanics. Group 3 included those most likely to use mechanics.

The sorting of majors into the three groups was done prior to administering the retest by consensus among the authors. Our consensus selections had to be modified based upon student performance for only two cases. Two majors initially placed in group 3 (mathematics and civil and environmental engineering) performed at a level more consistent with group 1. This resorting does not affect any quantitative conclusions of the paper, as will be demonstrated in Sec. III. The final list of majors for each group is shown in Table I.

The purest group from the perspective of investigating knowledge retention over a known interval is group 1, because members of this group generally experienced no significant review of the mechanics course content after finishing the freshman course apart from the electricity and magnetism course and vector calculus course that are required of all MIT students.

Table II summarizes the performance of the three major groups in the freshman course and on the senior retest. The performance of the different major groups in the freshman course was essentially equivalent. Standard analysis of variance (ANOVA) methods show no significant interaction of major group with either course grade or grade on the written analytic problems included on the freshman final exam. The interaction of major group with performance on the written analytic problems on the retest, however, was significant. The ANOVA result is $p = 3.7 \times 10^{-5}$ when all group 1 majors are included ($p = 4.2 \times 10^{-5}$ if we exclude mathematics majors and civil and environmental engineering majors from group 1). All three paired t-tests among the major groups showed significant variation in the average on the written analytic retest questions.

B. Construction of the exam

Our primary objectives in constructing the retest were to ensure accurate comparison of results with end-of-term freshman performance and also to place the focus of the exam on foundational topics that would be more likely to give a measurable retention score for all participants. Because all the written problems included on the 2005 final exam required recall of at least one topic taught in the last half of the course (rotation, oscillation, or orbits), we elected to construct the retest exam in part from more recent MIT final exams. This decision did not greatly compromise our ability to compare scores, since we did not have access to a full set of problem-by-problem scores for the 2005 final exams taken by the students as freshmen anyway. We did have scores for two of the problems on the 2005 exam, and we retained those two problems on the 2009 senior retest. The remaining written problems used for the final exam portion of the retest came directly from MIT exams for which we had full freshman class data (over 500 students). Thus, although the retest was easier than the final taken by the participants as freshmen in 2005, we had sufficient data to correct for this disparity as described in Appendix A.

The breakdown of the analytic questions by topic is shown in Table III. Note that on the 2005 final the students

TABLE I. Grouping of majors according to utilization of mechanics. *N* is the number of participants from each group.

	Included majors	*N*
Group 1	Biological Engineering, Biology, Brain and Cognitive Sciences, Civil and Environmental Engineering,[a] Literature, Management, Mathematics,[a] Political Science	26
Group 2	Chemical Engineering, Economics, Electrical Engineering and Computer Science, Materials Science and Engineering	21
Group 3	Aeronautics and Astronautics, Mechanical Engineering, Physics	9

[a]Initially in group 3; moved to group 1 based upon performance.

TABLE II. Major group average performance on their overall course grade, on the written questions in the fall 2005 final exam and on the written questions in the spring 2009 senior retest. The numbers in italics show the data for group 1 if mathematics majors and civil and environmental engineering majors are excluded.

Assessment	Group 1	Group 2	Group 3
Course grade	80(2)% *79(2)%*	82(1)%	82(3)%
Written problems fall 2005 final	55(4)% *52(4)%*	55(2)%	54(7)%
Written problems spring 2009 retest	27(3)% *25(3)%*	39(5)%	66(9)%

TABLE III. Comparison of the analytic questions on the 2005 final exam with those on the 2009 retest.

No.	2005 final[a]	2009 retest
1	collision with rotation	collision plus work or energy
2	orbit from spring force	collision with rotation
3	translation and rotation	Newton's 2nd law plus torque
4	orbit plus collision	orbit[b]
5	harmonic oscillation	harmonic oscillation[b]
6	gyroscopes	
7	torque and angular motion	

[a]Students were allowed to skip one of the seven questions.
[b]Students were allowed to skip either question 4 or question 5.

were allowed to skip one of the 7 questions. In order to maintain this element of choice in the retest while still ensuring that we obtained robust data on the more fundamental topics, we elected to specify that the students could skip either question 4 (covering orbits) or question 5 (covering oscillations) on the retest.

C. Analysis of gain and loss curves

The bulk of the quantitative analysis in this paper relies on the interpretation of gain and loss curves. In this section we give a brief introduction to the methods used.

Both gain and loss curves are plots of the score shift (score on retest minus score on an earlier test administration) versus the earlier test score. This is a *gain* curve (positive score shift) if learning has occurred between the administrations, and a *loss* curve (negative score shift) if forgetting has occurred.

The use of gain in physics education research was popularized by Hake [3], who showed that the gain on standardized tests of mechanics concepts is often well modeled by assuming that students in a given type of course learn, on average, a constant fraction of their maximum possible gain (defined by subtracting the student's pretest score from the maximum possible score on the test). If we assume that this average behavior is a good representation of the individual results of each student, then the curve defined by plotting the score shift of each student (posttest minus pretest) versus the pretest score for each student will be linear. The x intercept of the line will equal the maximum possible test score, and the absolute value of the slope will give one measure of the average *normalized gain* achieved by the group, where the normalized gain g is defined by

$$g = \frac{\text{posttest score} - \text{pretest score}}{100\% - \text{pretest score}}. \qquad (1)$$

A typical gain curve illustrating these features is shown in Fig. 1.

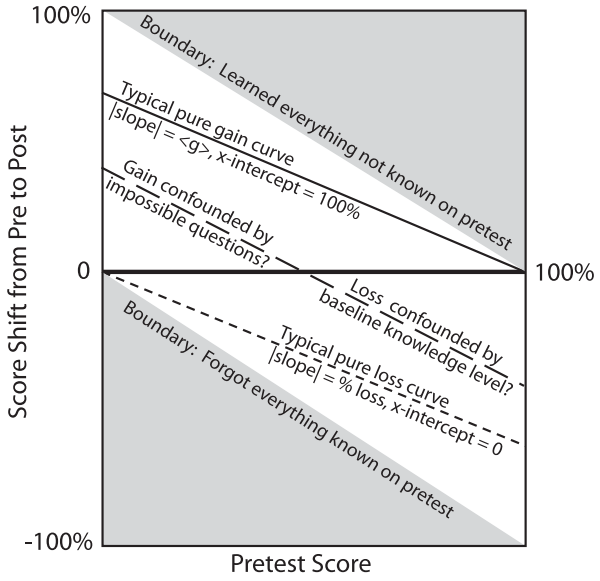

FIG. 1. The solid line represents a pure gain curve because it has its pretest score axis intercept at 100%. The absolute value of the slope of the gain curve is equal to the average normalized gain. The dotted line is a pure loss curve because it has its pretest score intercept at 0%. The absolute value of the slope of the loss curve is equal to the average fraction of knowledge lost. The dashed line is a sample linear fit to a data set which is neither pure loss nor pure gain.

In this work, we expected to see loss (negative score shift) rather than gain. Under the assumption commonly made in the literature [4,5] that the loss of knowledge over a retention interval is proportional to the initial knowledge, every student loses a standard fraction of their knowledge over the period of time between the pretest and the posttest. In this case, a plot of loss versus the pretest score will (like a gain curve with constant normalized gain) be linear with a negative slope. The absolute value of the slope equals the fraction lost. The only important difference between a loss curve and a gain curve is that the x intercept of the loss curve should be at zero rather than the maximum test score (it is impossible to lose knowledge if you know none of the items covered in the test initially, while it is impossible to gain knowledge if you already know everything covered in the test). An example loss curve illustrating these features is shown in Fig. 1.

We call a gain curve that decreases linearly to zero at 100% pretest score a case of *pure gain*, and one that decreases linearly from zero at 0% pretest score *pure loss*. Real data, however, may not be linear or may yield a linear fit with an x intercept that is neither 100% nor zero. Gain curves may exhibit an x intercept less than 100% if the test is difficult enough that even the best students cannot possibly master all the material. Loss curves may exhibit an x intercept greater than zero if the students have a store of essentially permanent baseline knowledge of the

material. Further, students may actually gain knowledge on some items in a given test while simultaneously losing knowledge on others. Thus, a real data set may generate an ambiguous curve similar to the middle line in Fig. 1.

III. RESULTS

A. 60% loss on analytic final exam problems among students not expected to review

We analyzed the distinct portions of the test independently, starting with the questions requiring written analytic responses. To allow for a comparison of the students' scores on the analytic portion of the 2005 final with their scores on the analytic portion of the 2009 retest, we assumed that the ability distribution of MIT freshmen is consistent from year to year. The validity of this assumption is supported by MIT admission data (available online [6]) as well as by student performance on the MBT administered at the beginning of the course as shown in Table IV. The sample of MBT pretest scores available to the authors includes all the freshman cohorts whose final exams were used to construct the retest.

The assumption of consistent ability distributions implies that we can generate a renormalized score on the freshman mechanics final taken in 2005 using the z scores achieved by our study participants as freshmen (z_{2005}). A z score is the deviation of the raw score achieved by the student from the class average on the exam divided by the standard deviation of the class on the exam. In order to renormalize the freshman scores, therefore, we first had to generate a mean (μ_{2009}) and standard deviation (σ_{2009}) for the senior retest using results from the administration of the questions to freshmen on their regular course final exams. The procedure used is explained in Appendix A.

A plot of the shift (score achieved on the analytic questions on the retest minus the renormalized score achieved on the analytic portion of the fall 2005 final exam) versus the renormalized fall 2005 analytic problem score is shown in Fig. 2. We have already shown in Sec. II A that there is a significant relationship between major group and retention, and this difference is evident in Fig. 2. The analysis of Sec. II A discovered a significant interaction between group number and score on the analytic portion of the retest while demonstrating that the interaction between group number and performance on the fall 2005 final exam is *not* significant (see Table II). This implies that the difference in performance on the retest arises from

TABLE IV. Performance of MIT 8.01 students on precourse administration of the MBT. Our participants were part of the 2005 cohort (shown in bold).

	2005	2007	2008	2009	Average
MBT	**14.85**	14.35	15.04	15.13	**14.84**
Pre	(**0.19**)	(0.19)	(0.18)	(0.16)	(**0.35**)

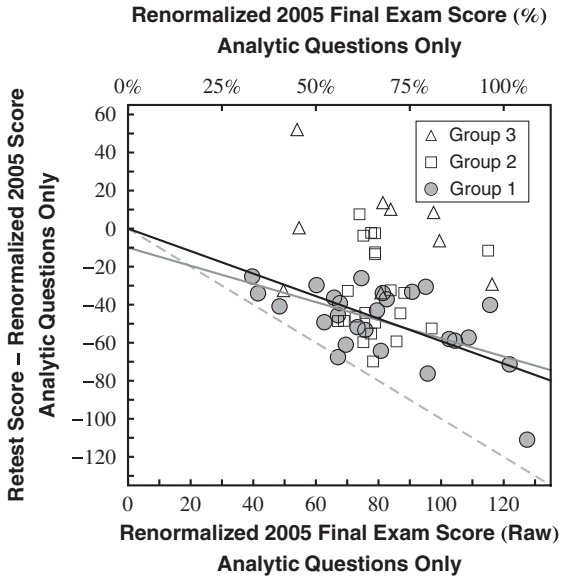

FIG. 2. Score shift between the end of freshman mechanics and the end of senior year on analytic problems versus freshman performance. The dotted line is the boundary corresponding to a score of zero on the retest. The solid lines are fits to the group 1 data only (gray with floating intercept, black with x intercept fixed at zero). The floating fit has x intercept -10.0 ± 11.9, which is consistent with zero. The slope of the fixed fit indicates knowledge loss of 0.59 ± 0.04 in the seven semesters since taking mechanics.

retention or review after the freshman course rather than performance in the freshman course.

The group 1 students, who were least likely to review the mechanics content in their course work, exhibit significant correlation between their fall 2005 score and their score shift ($r = -0.56$ for 26 students). The intercept of a linear fit to the group 1 data (Fig. 2) is consistent with zero, implying that the data are consistent with pure loss. With intercept fixed at zero, the group 1 students define a line with slope equal to -0.59 ± 0.04, meaning the students in group 1 lose 59% of the knowledge they had at the end of their freshman course over the following seven semesters at MIT. None of the group 1 students exceeded their freshman performance on the written analytic questions. The observed loss is essentially unchanged if the mathematics and civil and environmental engineering majors are removed from group 1. The slope of the best fit is identical and the correlation is still significant ($r = -0.45$ for 21 students).

The group 2 and group 3 students, who were thought likely to be exposed to some review of the concepts of introductory mechanics or at least to a reinforcement of the problem solving skills tested, do not exhibit any statistically significant trend. One of the 21 group 2 students and five of the nine group 3 students scored better on the retest than their renormalized freshman final exam score. We will revisit the group 3 data in Sec. IV.

TABLE V. Performance of seniors and (different) freshmen on five multiple-choice questions covering advanced topics. The scores are percentages, with 1σ errors in parentheses. Percent reduction is the reduction of the freshman average that would be required to yield the relevant senior average.

	Performance by group				
Question	1	2	3	All	Freshmen
Q1: Linear and angular acceleration of puck pulled by string.	37(9)	36(10)	78(14)	43(7)	73(2)
Q2: Internal forces conserve momentum.	17(8)	36(10)	44(17)	29(6)	63(3)
Q3: Angular momentum of a translating point particle.	50(10)	43(10)	56(17)	48(6)	66(3)
Q4: Period of mass-on-spring varies with square-root of mass.	42(10)	19(9)	33(16)	32(6)	63(3)
Q7[a]: Solid cylinder beats hollow cylinder down a ramp.	4(4)	5(5)	44(17)	11(4)	61(3)
Average	30(8)	28(7)	51(8)	32(6)	65(2)
% Reduction	**55(13)**	**58(10)**	**23(9)**	**51(9)**	

[a]Questions 5 and 6 were developed specifically for the retest so no comparison could be made.

B. 50% reduction on advanced concepts

The final exam portion of the retest contained seven multiple-choice conceptual questions dealing with advanced concepts like angular acceleration, angular momentum, and oscillations. The retest students did not answer questions of this type on their final exam in 2005, but five of the seven questions selected were taken from final exams given in the same course in more recent years. Table V shows that the seniors perform 50% worse than freshmen on these questions, though there is significant variation in the performance of the three groups. ANOVA yields $p = 0.01$ for equivalent averages among the groups on the questions listed in Table V. It is apparent from the data that group 3 stands out from groups 1 and 2, exhibiting greater knowledge of advanced concepts. The clearest separator is the significantly better performance of group 3 members on both the questions which involve torque (questions 1 and 7). The greater knowledge of group 3 and the approximately equal lesser knowledge of groups 1 and 2 is consistent with the attitudes of the group members to the advanced concepts covered in these multiple-choice questions, as discussed in Sec. V.

Because we do not have matched data for the retest participants as freshmen and as seniors, we cannot assert that the poorer performance of seniors relative to freshmen on the multiple-choice questions covering advanced concepts is truly a loss. We can report only that the seniors perform approximately 50% worse than the average freshman score on these questions. We have also not attempted to correct for random guessing, since this is challenging when dealing with questions for which some distractors are considerably more attractive to the students than others.

C. Gain and loss on the MBT

The mean score among the retest participants on the MBT was 17.6 ± 0.5 as seniors [7] versus 17.1 ± 0.5 as (postinstruction) freshmen. These scores give no indication of significant knowledge loss or gain. This is reflected in the score shift curve obtained for the MBT (Fig. 3). The curve is neither pure loss nor pure gain since it has an x intercept which clearly differs from zero and from 100%, even if we consider only students with majors in group 1.

These results suggest that we should examine the MBT data on a question-by-question basis (the full data set is presented in Appendix B). Doing so reveals evidence of significant *improvement* over the period from the end of freshman mechanics to the senior retest on five questions: 1, 2, 13, 19, 25. These questions are accompanied by four other closely related questions which most likely failed to demonstrate significant improvement due to saturation effects (the retest participants scored at or above 89% on each of these four). We therefore decided to analyze this group of nine questions as a separate subtest of the MBT.

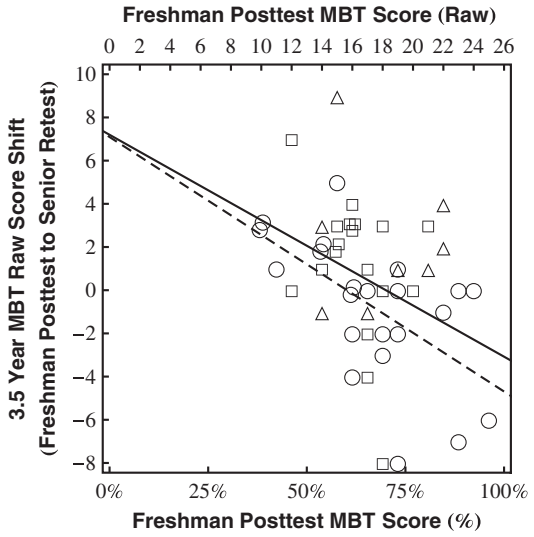

FIG. 3. Score shift from the end of freshman mechanics to graduation on the MBT versus freshman posttest score. The solid line is a fit to the data for all groups of students, the dashed line fits the group 1 students only. The x intercepts of the fits (70.0% and 60.0%, respectively) are clearly not consistent with zero or 100%.

TABLE VI. The MBT questions assigned to subtest G. Five of these questions (in bold) showed evidence of improvement by the seniors relative to their freshman posttest scores. The other four are directly related to the questions exhibiting gain, and each has a correct response rate above 79% on all three administrations of the MBT (freshman pretest, freshman posttest, and senior retest). Full question-by-question data are given in Table XI.

Subtest G of the MBT[a]	
Questions	Topic
1,**2**,3,23,24,**25**	Graphical kinematics
13,14	1D Equilibrium
19	2D Vector addition

[a]The 17 remaining questions were assigned to subtest L.

The content of this subtest, called subtest G (for ''gain'') is summarized in Table VI.

The gain curve for subtest G (Fig. 4) shows a very strong correlation between the students' score shifts on the nine-question subtest G over the 3.5 year retention interval (senior retest score minus freshman posttest score) and their original freshman posttest scores ($r = -0.81$ for 48 students), and is consistent with pure gain (the x intercept

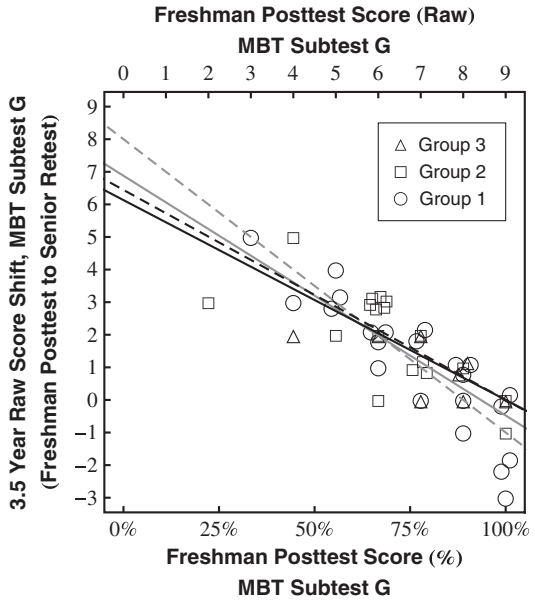

FIG. 4. Score shift from the end of freshman mechanics to graduation on subtest G of the MBT versus freshman posttest score on the same subtest. The solid lines are fits to the data for all groups (gray with floating intercept, black with x intercept fixed at 9). The dashed lines are fits to the group 1 data (gray with floating intercept, black with x intercept fixed at 9). The floating fit to the entire group has an x intercept of 8.42 ± 0.31, consistent with an x intercept of 9. The fixed fit to the entire group implies a normalized gain of 0.68 ± 0.05 over the 3.5 years since taking freshman mechanics.

8.42 ± 0.31 is consistent with 9, the maximum score on this subtest). The fit with intercept fixed at 9, shown in Fig. 4, implies an overall normalized gain of 0.68 ± 0.05 over the 3.5 years since taking freshman mechanics. This suggests that the material covered by the nine questions of subtest G is sufficiently ubiquitous in the MIT curriculum that all students, regardless of major, learn it well during their MIT careers.

We next consider the 17 questions which remain after separating out subtest G, which we will refer to as subtest L (for ''loss''). The questions included in this subtest cover a wide variety of topics including force, energy, impulse, and circular motion. Plotting the shift versus posttest curve for these 17 questions (Fig. 5) shows that removing the nine questions of subtest G resolves the ambiguity of the non-zero intercept in the loss versus posttest curve for the full MBT. The quality of the linear fit, however, is poor ($r = 0.27$) relative to the fits obtained for subtest G and for the loss data of group 1 on the analytic portion of the exam.

One possible explanation for the relatively poor fit obtained is the presence of baseline knowledge. It is reasonable to expect some baseline knowledge entering freshman mechanics among the participants, since MIT freshmen generally arrive with at least one year of high school physics. If we make the usual assumption [5] that preinstruction MBT scores measure baseline knowledge, it is possible to investigate how much of the knowledge *gained* during freshman physics remains at graduation. Plotting the score shift during the 3.5 years between freshman physics and graduation versus the gain made during the freshman course (Fig. 6) shows a stronger correlation than the gain versus freshman posttest score exhibited. The correlation for the group 1 data improves from $r = 0.27$ to $r = 0.55$, while that for the entire data set improves from

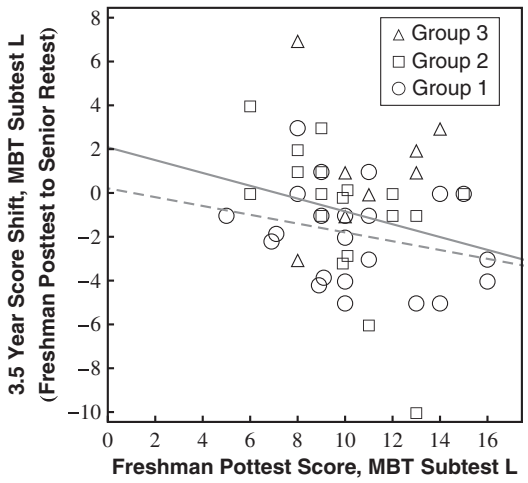

FIG. 5. Score shift from the end of freshman mechanics to graduation on subtest L of the MBT versus the score achieved on subtest L at the end of the freshman course. The solid line fits the data for all groups of students, the dashed line fits the group 1 data only.

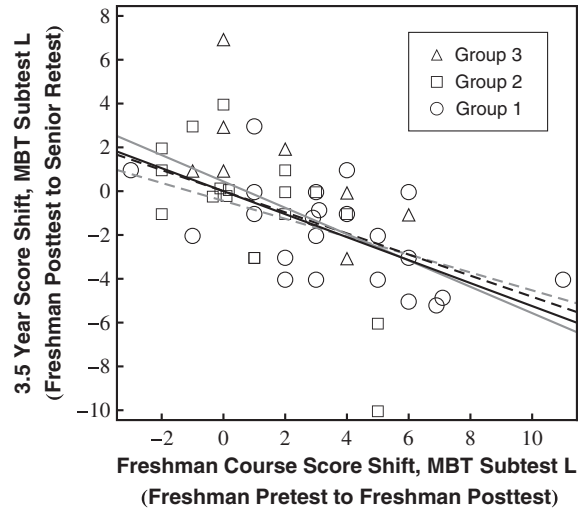

FIG. 6. Score shift from freshman posttest to senior retest on subtest L of the MBT versus shift during the freshman class. Solid fits are to the data for all groups (gray with floating intercept, black with x intercept fixed at zero). The fit with floating intercept has $p = 0.50$ for an intercept equal to zero. The dashed line fits the group 1 data only with x intercept fixed at zero. The slopes of the fits with fixed intercepts are -0.52 ± 0.09 and -0.48 ± 0.09, respectively.

$r = 0.26$ to $r = 0.59$. The intercepts retain their agreement with zero within errors.

To explore whether the loss from freshman posttest to senior retest is truly a function of the amount learned as freshmen (freshman posttest minus freshman pretest), we can perform a two-variable linear regression of the form

$$(s_{\text{retest}} - s_{\text{posttest}}) = \beta_{\text{post}} s_{\text{posttest}} - \beta_{\text{pre}} s_{\text{pretest}} + \alpha,$$

where s_{retest}, s_{posttest}, and s_{pretest} are the scores achieved on subtest L of the MBT in the senior retest, freshman posttest, and freshman pretest, respectively. The best-fit results are

$$\beta_{\text{post}} = -0.58 \pm 0.15, \qquad \beta_{\text{pre}} = -0.62 \pm 0.14,$$
$$\alpha = 0.05 \pm 1.5.$$

This fit explains 35.1% of the variance in the score loss, whereas assuming that the relevant variable is the amount learned as freshmen and performing a one-variable fit using the freshman score shift,

$$(s_{\text{retest}} - s_{\text{posttest}}) = \beta_{\text{shift}}(s_{\text{posttest}} - s_{\text{pretest}}) + \alpha,$$

yields the result

$$\beta_{\text{shift}} = -0.60 \pm 0.12, \qquad \alpha = 0.44 \pm 0.45,$$

and explains 35.0% of the variance in the score loss. Thus, the two-variable fit does not increase the explained variance and in fact results in slope parameters that are essentially indistinguishable from the single-variable fit. We take this as strong evidence that the amount learned during the freshman course (the freshman score shift) is the "natural" variable for removing baseline knowledge possessed by the students as entering freshmen and predicting subsequent loss on MBT subtest L.

Interpreting Fig. 6 as pure *loss of gain* and constructing fits to the data with the intercept fixed at zero gives a best-fit slope of -0.48 ± 0.09 for the group 1 data and -0.52 ± 0.09 for the entire data set. Thus, we conclude that on the 17 questions making up subtest L of the MBT, all students lost approximately half of the knowledge they *gained in the course*. Moreover, there appears to be a baseline of long-term knowledge relevant to this portion of the MBT among the students in our sample. Given that the mean score of the retest participants on subtest L was 8.1 ± 0.4 preinstruction, 10.4 ± 0.4 postinstruction, and 9.4 ± 0.5 as seniors, it is reasonable to characterize the baseline knowledge as substantially larger than the amount of learning or forgetting observed on this subtest during the participants' time at MIT.

IV. MOTIVATION

From the beginning, we were concerned that lack of motivation in taking the senior retest could create the appearance of skill loss. Our incentive payments were designed to encourage motivation. To check the motivation level that was achieved, we sent an Email containing a five-question survey to participants along with a summary of the research findings and notification of disbursal of their payment. The Email was sent 50 days after the retest was administered and contained the question, "How would you rate your effort level on the retest? (If you feel your effort level changed during the course of the test, and you remember what you were working on at the time of that change, please specify when the change occurred.)" Eleven out of the 56 participants Emailed responses.

Ten of the 11 responses explicitly indicated high effort either on all questions or on all multiple-choice questions. The 11th did not explicitly indicate high effort on multiple-choice questions, but also did not preclude it [8]. We therefore consider it unlikely that motivation had a significant influence on the results on the MBT or the multiple-choice questions covering advanced concepts.

By contrast, the responses were approximately evenly divided with regard to the written analytic problems. Five of the 11 students expressed a low or at least reduced motivation on these problems. Because our analysis of the analytic problems focused on group 1, it is particularly important to examine the responses from this group. Five of the 11 respondents had majors in group 1. Two of these five reported full motivation on the analytic problems while three reported less than full motivation. These data are limited, but they support the claim that our group 1 sample contained some individuals who exerted full effort on the test and also some who did not. Proceeding upon this assumption, we can check for any signal of a split in the

knowledge loss exhibited by students that might indicate a significant performance difference between those who exerted full effort and those who did not.

A histogram of the loss fraction on the analytic problems of the retest for group 1 is shown as Fig. 7. This histogram shows no evidence of a split in retention levels. There is no evidence that high motivation on the retest would eliminate or even dramatically reduce the loss in performance observed on the analytic questions of the final exam among group 1 students as seniors relative to freshmen. In fact, one of the group 1 students who indicated high motivation in their survey response qualified it with, ''I tried pretty hard, but there were some fundamental concepts that I did not remember, and without them, I knew I couldn't figure out the problem. If I felt like I could reason out the solution, I would have.''

Motivation on the final exam taken as freshmen is also relevant to our results. Many instructors can give anecdotal reports of students ''gaming the system'' by precomputing the final exam score needed to achieve some desired course grade and then adjusting their effort on the final exam to the level appropriate (from the student's perspective) to achieve that target score. If any of the students participating in our study employed this approach as freshmen, it would constitute a source of bias in our results.

One way to assess student motivation on the freshman final is to assume that the z score of students (their rank in their peer group) tends to remain stable. Computing the deviation of the final exam z score achieved from the average z score on midterm exams gives one possible measure of motivation. A drop in z score on the final is potentially a signal of low motivation while an anomalously high z score could indicate strong motivation.

We have correlated the measure $z_{\text{final}} - \langle z \rangle_{\text{midterms}}$ with the observed loss of knowledge on the retest. We expect

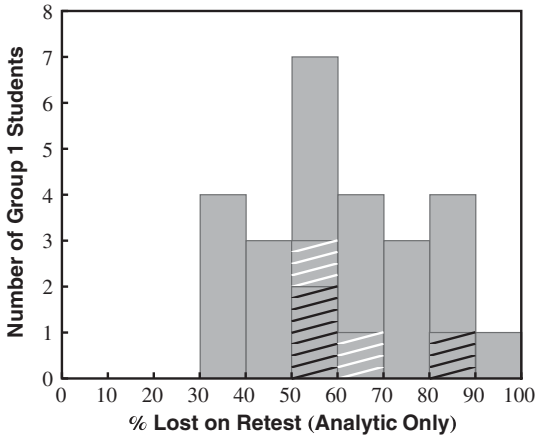

FIG. 7. Histogram showing the number of students in group 1 with a given loss fraction on the analytic problems of the retest. The hatched areas are a stacked histogram of the students who self-reported full (white) or less than full (black) effort on this portion of the retest.

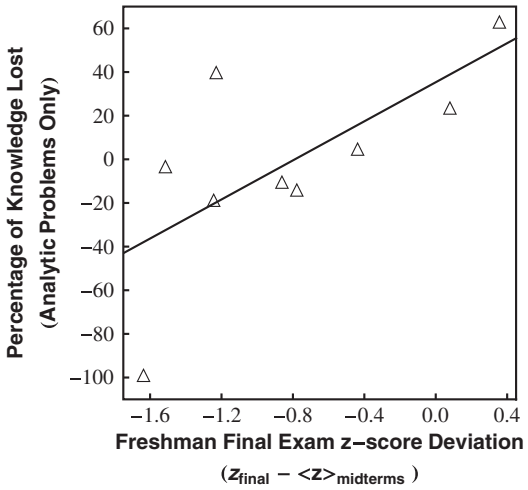

FIG. 8. Plot of percentage loss on analytic questions versus deviation of final exam z score from mean midterm z score for group 3 students. The y intercept of the linear fit is 35% ± 20%.

that students who overperformed on the freshman final would appear to have extra loss as seniors, while students who underperform on the freshman final would seem to show less loss. The observed correlation is weakly negative for groups 1 and 2, which contradicts the expected relationship; these correlations are not significant, however. For group 3, on the other hand, the correlation is significant and positive. The correlation between the z score deviation on the freshman final exam with percentage of knowledge lost on the retest is +0.68, implying that anomalous performance on the final accounts for about 46% of the variance in retention observed among group 3 students. The y intercept of a linear fit to the group 3 loss fraction versus z-score shift data (Fig. 8) is 35% ± 20%. This intercept could be taken to represent a measure of the loss fraction of group 3 after correction for anomalous performance on the freshman final exam.

Group 3 was not the only group to show significant levels of underperformance on the freshman final exam. The average deviation of final exam z score from mean midterm z score for group 3 was -0.81 ± 0.23, but group 2 actually exhibited a stronger average deviation of -0.99 ± 0.11. Anomalous final exam scores among group 2 members, however, have essentially zero correlation with the loss they exhibit on the retest ($r = -0.04$ for 21 students).

TABLE VII. CLASS categories [1] exhibiting significant shifts over four years. The 2005 data are preinstruction. Significant shifts are shown in bold.

Category	% Favorable 2005	% Favorable 2009	% Unfavorable 2005	% Unfavorable 2009
Personal interest	51(4)	50(4)	**17(3)**	**27(4)**
Real world connection	**52(4)**	**68(3)**	18(3)	20(3)
Sense making/effort	68(4)	65(3)	**8(2)**	**18(3)**

TABLE VIII. Responses of the retest participants to CLASS statements 14 and 30 as preinstruction freshmen (2005) and graduating seniors (2009). Significant shifts are in bold.

Statement	% Favorable 2005	% Favorable 2009	% Unfavorable 2005	% Unfavorable 2009
I study physics to learn knowledge that will be useful in my life outside of school.	**38(7)**	**20(5)**	**23(6)**	**44(7)**
Reasoning skills used to understand physics can be helpful to me in my everyday life.	**58(7)**	**85(5)**	8(4)	5(3)

TABLE IX. Responses to questions on our demographic survey. Course 8.01 is freshman mechanics.

Question	% Yes	% No
Do you think the material taught in 8.01 will be useful to you after graduating?	54(7)	46(7)
Do you feel 8.01 should be a required course for students in your major?	93(3)	7(3)

Group 1, the group of principal interest, had the smallest average deviation (-0.39 ± 0.17) and a nonsignificant correlation with loss ($r = -0.17$ for 26 students).

In conclusion, we find no evidence that motivation significantly biases the results presented in Sec. III. It appears that the vast majority of retest participants exerted full effort on the multiple-choice instruments, that any variation of effort among group 1 students on the written analytic problems was overwhelmed by the extent of their knowledge loss, and that group 1 students did not tend to "game the system" on the freshman final exam.

V. STUDENT ATTITUDES

Student attitudes were measured by the CLASS standard instrument [1] and also by questions on a demographic survey generated by us. The CLASS was previously administered to the students as they entered their freshman course in 2005. Three of the nine categories commonly used to analyze CLASS data exhibited significant shifts (Table VII).

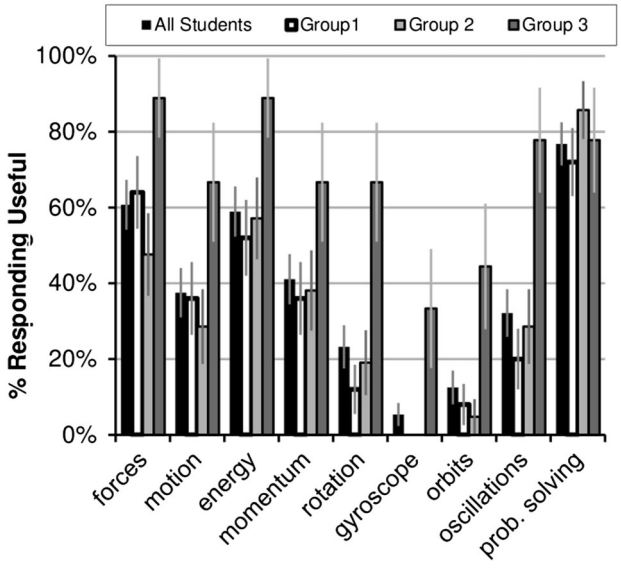

FIG. 9. Frequency with which various topics taught in mechanics were circled as "useful to you since taking [mechanics]" by students on the demographic survey.

Curiously, the personal interest category shifts toward unfavorable responses while the real world connection shifts toward favorable. Looking at the statements making up these categories, we find that the students draw a distinction between the factual content of the mechanics course and the general reasoning skills that are taught. Both the CLASS and our own survey suggest that the students find the general skills more valuable than the factual content. The evolution of the responses to CLASS statements 14 and 30 (Table VIII) indicates that this distinction becomes more pronounced during their four years of undergraduate education. By the end of their four years at MIT, almost half of the students in our study display unfavorable attitudes about the relevance of the material taught in the introductory physics course to their own lives after graduation (Tables VIII and IX). By contrast, about three-quarters of the seniors in our study indicate that the problem solving taught in introductory physics was useful beyond the scope of that course (Fig. 9), 85% of them now feel that the reasoning skills taught in physics are valuable in everyday life (Table VIII), and over 90% believe that physics should retain its status as a required course in the MIT curriculum (Table IX).

VI. CONCLUSIONS

A. Quantitative findings

Table X presents a recap of the key quantitative outcomes of our study.

We summarize this work from the perspective of what the various groups lost or gained on the three main categories of physics problems.

- Group 1 (major unrelated to physical science) lost ~55% on analytic problems, advanced physics concepts, and simpler physics concepts on the MBT (with MBT loss measured against knowledge *learned* in the freshman course rather than absolute knowledge).

TABLE X. Summary of percentage loss (−) or gain (+) for each major group in each data set studied.

Summary of data sets		% Loss (−) or gain (+) by group			
Type of material	Gain or loss relative to	1	2	3	All
Analytic problems	Renormalized freshman final Exam score	−59(4)	−41(7)	−3(13)	−44(4)
Graphical kinematics, equilibrium, vector addition (MBT)	Knowledge lacked on freshman posttest	+72(9)	+70(8)	+49(12)	+68(5)
Basic physics concepts, e.g., acceleration, force, energy (MBT)	Freshman course gain (posttest − pretest)	−48(9)	−61(23)[a]	−20(44)	−52(9)
Advanced physics concepts, e.g., rotation, oscillation, orbits	Freshman average[b] performance	−55(13)	−58(10)	−23(9)	−51(9)

[a] −95(25) if the conspicuous outlier at the point (5, −10) is included in the fit.
[b] Lacking data for the retest participants as freshmen, we can only report reduction relative to freshman average performance.

- Group 2 (major not involving mechanics) also lost ∼55% on advanced physics concepts and simpler physics concepts on the MBT. Their performance on the analytic questions showed wide variation, with some indication of a high-retention and a low-retention subgroup.
- Group 3 (major directly involving mechanics) neither lost nor gained significantly on analytic problems and lost only 20%–25% on advanced physics concepts and simpler concepts covered on the MBT.

Thus, groups 1 and 2 lose ∼55% across the board except that group 2 has possibly lost less (∼ 40% on average) on analytic problems (consistent with the interpretation that group 2 majors demand analytic prowess, but not in the domain of mechanics). It is reassuring that group 3 retains all their ability on analytic mechanics problems. It is, however, somewhat troubling that they do exhibit loss on conceptual problems. This suggests that they do not review or build on the conceptual foundations of mechanics to the degree that might be expected in more advanced courses. This is consistent with the finding of Pollack [9] that upper-division electricity and magnetism (E&M) courses do not affect performance on conceptual questions given in the freshman E&M course.

Taken together, these results conform to the adage "use it or lose it." They are a strong argument for spiral learning, the return to a topic learned previously in a subsequent course.

We now turn to subtest G of the MBT, on which all three groups exhibited substantial normalized gains relative to their freshman year posttest. This subtest concerns general skills like understanding graphs, vectors, and the calculus of kinematics. The exact source of this further learning is unclear. Certainly, all MIT students can be expected to review vectors in the required E&M course and the calculus sequence, but these are also generally freshman or sophomore level courses leaving three years' opportunity for forgetting. It may be that these particular questions rely on skills that are so basic (e.g., reading and comprehending graphs) that they are reviewed throughout the curriculum in many majors.

Our study is consistent with, yet calls into question, previous studies of retention on concept tests. Many studies of retention on other standardized instruments testing conceptual physics knowledge [9–12] show little or no loss of knowledge over retention intervals ranging from several months to several years. Indeed, this is consistent with our finding that the average score on the MBT among the retest participants is remarkably stable over the interval from the completion of freshman mechanics to graduation. However, our study investigates in more detail and reveals that this overall stability is the net result of a substantial gain on general mathematical skills counterbalanced by a simultaneous loss in items involving core physics topics, on which the loss is comparable to the loss on analytic items *if* we control for the baseline knowledge possessed by the students before taking freshman mechanics. Therefore, we strongly recommend that future studies of retention on conceptual instruments investigate whether the loss or gain is uniform over the entire body of material tested or if it is the result of simultaneous gain and loss in a manner similar to the MBT results presented here, and we stress the importance of gauging loss relative to the amount learned rather than to the amount known.

B. Attitudes and motivation

Survey responses indicate that during their undergraduate education student attitudes evolve to increasingly value the reasoning and problem solving skills taught in physics while simultaneously placing lesser value on mechanics concepts as relevant to their everyday lives. This demonstrates the alignment of the calls in the physics education research community for a greater emphasis on developing problem solving expertise even at the expense of some factual or procedural content with the needs of students (see, e.g., [13–16]).

The data on student motivation are less clear and point to interesting avenues of research. Our study suggests that group 1 majors have lost so much knowledge that

motivation on the retest was essentially irrelevant. We see evidence that group 3 majors game the system on their freshman final, frequently underperforming relative to their knowledge, and that this effect might be significant enough to explain away any gains seen in their knowledge on the retest. Group 2 students also appear to underperform on the freshman final relative to their performance earlier in the class, but this does not seem to be a gaming of the system (their retention is uncorrelated to their underperformance); it could simply be an indication that the material at the end of the MIT course is significantly harder for these students to learn than the earlier topics.

ACKNOWLEDGMENTS

The authors gratefully acknowledge R. Romano from the Registrar's Office and A. Clark from the Office of the Dean for Undergraduate Education for assisting with the logistics of the retest study. This work was supported by the National Science Foundation under Grants No. PHY-0757931 and No. DUE-1044294.

APPENDIX A: RENORMALIZATION OF ANALYTIC SCORES

The procedure for generating the assumed mean for the 2009 retest can be illustrated in generality by assuming the test was composed of N problems, each administered on the regular MIT final exam in different years. Using the data from these administrations, we were able to find question-by-question averages (μ_i, $i = 1, 2, \ldots, N$) and standard deviations (σ_i). The expected average of MIT freshmen on the 2009 test was then calculated as

$$\mu_{2009} = \sum_{i=1}^{N} \mu_i.$$

Finding the expected standard deviation of the entire examination in terms of the standard deviation of the problems is not as simple as estimating the mean. If the deviations of a given student on each problem are completely random, then we would expect the standard deviations of the problems to add in quadrature:

$$\sigma^2_{\text{uncorrelated}} = \sum_{i=1}^{N} \sigma_i^2.$$

If, on the other hand, the deviation of a given student on each problem is perfectly correlated and the variance of each problem is identical, the expected relationship is

$$\sigma^2_{\text{perfect correlation}} = N \sum_{i=1}^{N} \sigma_i^2.$$

In the real world, we would expect that neither extreme is realized. For this reason, we took an empirical approach to estimating the ratio of σ^2_{2009} to the sum of the squares of the individual standard deviations.

We used test data from several MIT final exams and constructed 35 different four-problem "tests" by choosing different problem combinations from each final. Finding the actual standard deviation of these four-problem tests and comparing to the sum of the standard deviation of the individual problems (Fig. 10) leads us to estimate that on MIT final exams

$$\sigma^2_{\text{4-problem test}} = (1.88 \pm 0.33)\left(\sum_{i=1}^{4} \sigma_i^2\right). \tag{A1}$$

The value 1.88 is comfortably between the limits of 1.0 (no correlation of a given student's performance on the four problems) and 4.0 (complete correlation of performance). This empirical relationship implies

$$\sigma_{2009} = 17.5\% \pm 1.6\%, \tag{A2}$$

where we have expressed the standard deviation as a percentage of the total points available on the retest. The value 17.5% corresponds to 21 points out of a total 120 possible on the retest. (For comparison, the standard deviation calculated from the scores of the retest participants is 29 points or 24%. This higher value is clearly influenced by the observed variation in retention.)

Once we had an estimated class mean and standard deviation for the 2009 retest, the renormalized score of the students on their freshman final ($\bar{s}_{2005}$) was calculated using

$$\bar{s}_{2005} = z_{2005}\sigma_{2009} + \mu_{2009}. \tag{A3}$$

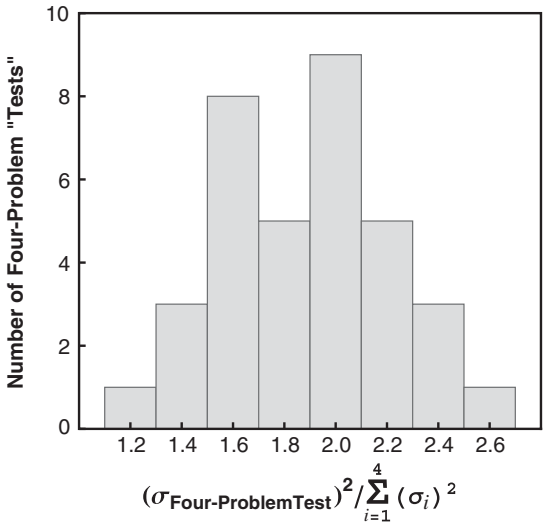

FIG. 10. Histogram showing the distribution of the ratio of the square of the standard deviation observed in 35 four-problem "tests" (combinations of four MIT final exam problems) to the sum of the squared standard deviations of the four individual problems.

Because the retest had a higher mean score than the 2005 exam, some of the renormalized scores exceeded 100% (120 points).

APPENDIX B: DETAILED MBT RESULTS

Table XI presents the full set of MBT data for the retest participants. This includes pretest and posttest administrations during the fall 2005 mechanics course and our retest administration in spring 2009. The table presents the percentage of retest participants correctly answering each question on the MBT during the fall 2005 pretest (''Pre'') the fall 2005 posttest (''Post'') and the 2009 senior retest (''Retest'').

Three questions on the MBT (10, 11, and 26) show significant loss on the retest relative to the posttest. Two of these (10 and 11) test understanding of conservation of energy (specifically the transformation of gravitational energy to kinetic). The final question (26) tests understanding of the acceleration experienced by an object in near Earth free fall.

TABLE XI. Detailed MBT results. Questions included in subtest G are denoted by a (G) following the number. Uncertainties are given in parentheses. Bold numbers denote a significant posttest to retest shift.

	% Correct				% Correct		
No.	Pre	Post	Retest	No.	Pre	Post	Retest
1 (G)	75(7)	69(8)	**94(4)**	14 (G)	81(6)	83(6)	90(5)
2 (G)	81(6)	69(8)	**90(5)**	15	58(9)	79(7)	81(6)
3 (G)	81(6)	88(5)	98(2)	16	50(10)	63(9)	60(9)
4	58(9)	46(10)	56(10)	17	65(9)	65(9)	75(7)
5	17(6)	40(9)	21(7)	18	23(7)	21(7)	27(8)
6	83(6)	92(4)	90(5)	19 (G)	52(10)	65(9)	**90(5)**
7	48(10)	48(10)	52(10)	20	33(8)	48(10)	48(10)
8	52(10)	79(7)	67(8)	21	73(8)	96(3)	88(5)
9	31(8)	52(10)	44(10)	22	52(10)	44(10)	44(10)
10	67(8)	92(4)	**75(7)**	23 (G)	85(6)	79(7)	92(4)
11	33(8)	75(7)	**48(10)**	24 (G)	90(5)	92(4)	88(5)
12	13(5)	25(7)	23(7)	25 (G)	71(8)	63(9)	**85(6)**
13 (G)	60(9)	65(9)	**85(6)**	26	50(10)	77(7)	**46(10)**

[1] W. K. Adams, K. K. Perkins, N. S. Podolefsky, M. Dubson, N. D. Finkelstein, and C. E. Wieman, New instrument for measuring student beliefs about physics and learning physics: The Colorado Learning Attitudes about Science Survey, Phys. Rev. ST Phys. Educ. Res. **2**, 010101 (2006).

[2] D. Hestenes and M. Wells, A mechanics baseline test, Phys. Teach. **30**, 159 (1992).

[3] 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).

[4] E. J. F. M. Custers, Long-term retention of basic science knowledge: A review study, Adv. Health Sci. Educ. **15**, 109 (2010).

[5] G. B. Semb and J. A. Ellis, Knowledge taught in school: What is remembered?, Rev. Educ. Res. **64**, 253 (1994).

[6] Data available online from MIT's Office of the Provost, http://web.mit.edu/ir/cds/index.html, retrieved February, 2011.

[7] The nature of our study (a retest using the same instruments as the students used as freshmen) forces us to use the MBT again, even though it has not been validated for seniors (as opposed to freshmen or high school students). In fact, one of our main conclusions suggests that it may not be appropriate for scoring MIT seniors because subtest G of the MBT is too easy to provide significant discrimination.

[8] The exact response was ''low effort level—gave up on problems I might have kept working on as a freshman.''

[9] S. J. Pollock, Longitudinal study of student conceptual understanding in electricity and magnetism, Phys. Rev. ST Phys. Educ. Res. **5**, 020110 (2009).

[10] G. Francis, J. Adams, and E. Noonan, Do they stay fixed?, Phys. Teach. **36**, 488 (1998).

[11] Y. J. Dori, E. Hult, L. Breslow, and J. W. Belcher, How much have they retained? Making unseen concepts seen in a freshman electromagnetism course at MIT, J. Sci. Educ. Technol. **16**, 299 (2007).

[12] M. A. Kohlmyer, M. D. Caballero, 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).

[13] L. C. McDermott, Millikan Lecture 1990: What we teach and what is learned—Closing the gap, Am. J. Phys. **59**, 301 (1991).

[14] P. W. Laws, Millikan Lecture 1996: Promoting active learning based on physics education research in introductory physics courses, Am. J. Phys. **65**, 14 (1997).

[15] E. F. Redish, Millikan Lecture 1998: Building a science of teaching physics, Am. J. Phys. **67**, 562 (1999).

[16] A. Van Heuvelen, Millikan Lecture 1999: The workplace, student minds and physics learning systems, Am. J. Phys. **69**, 1139 (2001).