# Measuring the Effect of Formative Assessment Techniques in Physics at East Stroudsburg University

Robert A. Cohen
*Department of Physics, East Stroudsburg University,*
*East Stroudsburg, Pennsylvania*

Review of Content Study carried out as part of the CETP-PA Project

Previous results of this study were submitted to
the *CETP-PA Monograph*, October, 2005

*Author address:* Dr. Robert A. Cohen, Department of Physics,
East Stroudsburg University, East Stroudsburg, PA 18301.
E-mail: rcohen@po-box.esu.edu

April 23, 2007

**Abstract**

Three formative assessment techniques were implemented in several sections of PHYS131 (Fundamental Physics I) at East Stroudsburg University and the performance of those students were compared to sections in which the techniques were not implemented. It was found that the techniques had a positive impact on student learning.

| Section | Semester | Methodology | Textbook | Instructor |
|---|---|---|---|---|
| F1 | Fall-2002 | Remotes, email and text | Cohen | tenured |
| F2 | Fall-2003 | Remotes, email and text | Cohen | tenured |
| F3 | Fall-2004 | Remotes, email and text | Cohen | tenured |
| F4 | Fall-2005 | Remotes and text | Cohen | tenured |
| F5 | Fall-2006 | Remotes, email and text | Cohen | tenured |
| C1 | Fall-2003 | Traditional | Cutnell/Johnson | one-year temporary |
| C2 | Spr-2004 | Traditional | Cutnell/Johnson | one-year temporary |
| C3 | Sum-2004 | Traditional | Cutnell/Johnson | one-year temporary |
| T1 | Fall-2006 | Traditional | Cutnell/Johnson | tenured |
| T2 | Fall-2002 | Traditional | Urone | tenured, highly regarded |
| T3 | Fall-2004 | Text | Cohen | tenured, highly regarded |
| T4 | Spring-2006 | Traditional | Knight | tenured |

**Table 1:** A summary of the twelve different sections utilized for this study.

## 1. Context

Three techniques, all designed to improve formative assessment, were implemented in selected sections of a physics course (Fundamental Physics I) at East Stroudsburg University. The course is offered every semester and is part of a two-semester sequence taken by life science majors (e.g., biology, pre-pharmacy, etc.), usually in their junior year.

In the study, twelve sections were examined, of which eleven were of Fundamental Physics I and one which was the calculus-based version of the same course. These will be referred to as either "control" sections (indicated as C1 through C3), "formative assessment" sections (indicated as F1 through F5) or "test" sections (indicated as T1 through T4). All sections identified content understanding as measured by an in-house multiple-choice survey (see section 2).

The twelve sections are summarized in Table 1.

### *a. Formative assessment sections*

The five formative assessment sections (F1-F5) utilized the following three formative assessment techniques:

1. Students were required to buy infrared response pads (eInstruction Corp.) that were used during class to assess student understanding and guide instruction. To build consensus, students were encouraged to talk to their peers and explain the rationale for their answers, particularly for those questions where an initial polling did not reveal a consensus. Additional time was then spent on those areas that produced lack of a consensus or low success. This activity was designed to do the following:

   (a) Allow students to get a sense of where they stood relative to the rest of the class.

   (b) Provide the instructor with a sense of how well the students were meeting the lesson objectives and thus allow the instructor to adapt instruction to student

needs.

(c) By putting into words the reasons for their choices, students were forced to examine the extent of their understanding (vs. regurgitation of facts and figures).

2. Students were required to email the instructor before each class with questions about the readings. The lessons were then designed to address those questions. This activity was designed to do the following:

   (a) Force students to focus on the extent of their understanding while reading the material.

   (b) Provide the instructor with insight into the areas students were experiencing difficulty so that the instructor could address those areas during class time. This aspect is similar to the Just-in-Time Teaching (JiTT) method discussed by Novak et al. (1999).

3. An innovative textbook (Cohen, 2006) was developed that encouraged formative assessment by embedding rhetorical questions and homework questions within the text. In addition, the textbook reordered the sequence of topics to clarify how each new concept addressed weaknesses in previous concepts. Developed in-house, the textbook was printed in black and white (no color pictures) with few examples and no pictures (only line drawings) or ancillary materials. The textbook was revised to correct errors and clarify concepts after each usage so different sections used different versions of the textbook (compare Cohen, 2006, with Cohen, 2002, 2003, 2004 and 2005).

The formative assessment sections (F1-F5) were taught each fall by the author of the study. The author holds a tenured position in the physics department, is familiar with formative assessment techniques and holds PA teaching certification in physics. All five of these sections involved the three assessment techniques discussed above to various degrees except for test section F4, which did not use the email technique (technique #2 in the list). Other than the formative assessment techniques, the instruction mainly consisted of lecture, demonstrations and discussion/questioning.[1]

The reason why section F4 did not utilize the email technique was to identify whether the emailing activity was crucial. Such an activity required a great deal of time on the instructor's part to respond to each question. As the most demanding activity, it seemed to be the first element that could be dropped by a non-dedicated instructor.

*b. Control sections*

The three "control" sections (C1, C2 and C3) were taught by a one-year temporary instructor, using none of the three formative assessment techniques. Rather, instruction was limited to the "traditional" methods of lecture and demonstrations, with an emphasis on engineering applications, and no collaborative or group activities. The instructor had a masters degree in engineering and PA teaching certification in physics, but little prior experience teaching a college-level physics class. A popular commercially-available textbook (Cutnell and Johnson, 2004) was used as the required text.

[1] Sections F1 and F2 occasionally used group activities also.

*c. Test sections*

The four test sections "tested" various hypotheses about the effect of different methodologies other than the three formative assessment techniques utilized in the "formative assessment" sections.

All four test sections were taught by tenured instuctors and utilized "traditional" methods of instruction.[2] Test section #1 used the same textbook as the control sections but a different homework structure.[3] Section T2 used a a commercially-available textbook (Urone, 2001) that is somewhat less popular than the one by Cutnell and Johnson (2004). Section T3 used the innovative textbook (formative assessment #3 above). Section T4 used a calculus-based textbook (Knight, 2004).

Of the four sections, one (T4) was taught by same instructor as the formative assessment sections (F1-F5). Two of the sections (T2 and T3) were taught by an instructor recognized as an exemplary physics instructor by both students and faculty (as evidenced by peer evaluations[4] and student evaluations).

## 2. Measurement of Student Learning

Student performance was measured by a 17-question multiple-choice survey. Questions on the survey were taken from the Force Concept Inventory (Hestenes et al, 1995), a well-tested instrument for measuring conceptual understanding of forces, and supplemented by our own questions on graphing, kinematics and vectors. Questions were selected from those that students typically are expected to answer correctly (i.e., they cover seemingly basic ideas of physics) but don't. The survey was also constructed with an eye toward making the survey as short as possible to simplify its administration. Each class received the instrument twice, at the beginning of the semester and at the end.

While all sections gave the instrument as a non-graded activity the first day of class, the implementation of the instrument at the end of class varied from section to section. In most sections, the questions were split up and administered as part of the exams. In some sections, however, the survey was again given as a non-graded activity, albeit near the last day of class. To determine if the method of implementation affected the results, both techniques were used in one of the sections. It was found that while scores on individual questions varied, the overall scores and the conclusions reached by examining the overall scores were not significantly dependent on how the instrument was administered.

The actual survey had twenty questions. However, only the first 17 were used for this study. Question 18 was designed to assess whether students understood the difference between laws and theories. Since the distinction is not typically discussed in this type of

---

[2] Traditional methods include lecture, demonstrations and questioning, with no collaborative or group activities.

[3] Students were assigned multiple-choice questions. To get credit, they had to not only choose the correct answer but also provide the correct reasoning.

[4] According to peer evaluations, the main instructional difference between section T2 and the control sections (C1-C3) was that the T2 instructor utilized questioning more effectively and incorporated explanations that were more accessible to the students.

| Section | Semester | N | Avg pre-test score | Avg post-test score | Change |
|---|---|---|---|---|---|
| F1 | Fall-2002 | 27 | 33% | 54% | +21% |
| F2 | Fall-2003 | 17 | 29% | 56% | +27% |
| F3 | Fall-2004 | 26 | 31% | 68% | +37% |
| F4 | Fall-2005 | 28 | 27% | 62% | +35% |
| F5 | Fall-2006 | 24 | 29% | 61% | +32% |
| C1 | Fall-2003 | 8 | 31% | 33% | +2% |
| C2 | Spr-2004 | 16 | 32% | 34% | +2% |
| C3 | Sum-2004 | 5 | 26% | 28% | +2% |
| T1 | Fall-2006 | 29 | 26% | 69% | +43% |
| T2 | Fall-2002 | 20 | 29% | 53% | +24% |
| T3 | Fall-2004 | 39 | 28% | 50% | +22% |
| T4 | Spring-2006 | 17 | 37% | 70% | +33% |

**Table 2:** A summary of the pre/post scores for the twelve sections utilized for this study.

physics class, it was not used in the comparisons. Questions 19 and 20 diagnose hypothetico-deductive thinking and were selected from the Classroom Test of Scientific Reasoning (revised edition, August 2000) by Anton E. Lawson at Arizona State University (Lawson, 1978). They were included only to identify whether performance on the survey or in the class was related to performance on this pair of questions (no relationship was observed).

## 3. Results

### *a. Pre and post scores*

The results of the study are shown in Table a and the left graph of Figure 1.

The formative assessment sections all experienced at least a 20% gain in scores from pre to post. In addition, there was a steady improvement in scores during the first three years of the study before leveling off. Possible reasons for the variation in scores among sections F1 through F5 are discussed in section 4.

In comparison, scores in sections C1, C2 and C3 (taught by the one-year temporary instructor using traditional techniques) remained relatively the same for all three sections (an increase of +2% pre vs. post), reflecting minimal growth in concept understanding. This also implies that experienced instructors can make a significant difference in student learning.

Of the four test sections (T1-T4), two sections (T1, T4) experienced scores similar to the highest of the formative assessment sections (F3) while two sections (T2, T3) experienced scores similar to the lowest of the formative assessment sections (F1). As such, a closer examination of the test section scores is warranted.

The biggest gain was obtained by section T1 (traditional instructional techniques and textbook). However, the post-test was provided as part of a take-home exam. Consequently, not only did students have a longer period to complete the questions but they also had access to the textbook and notes. In addition, comparison between individual students

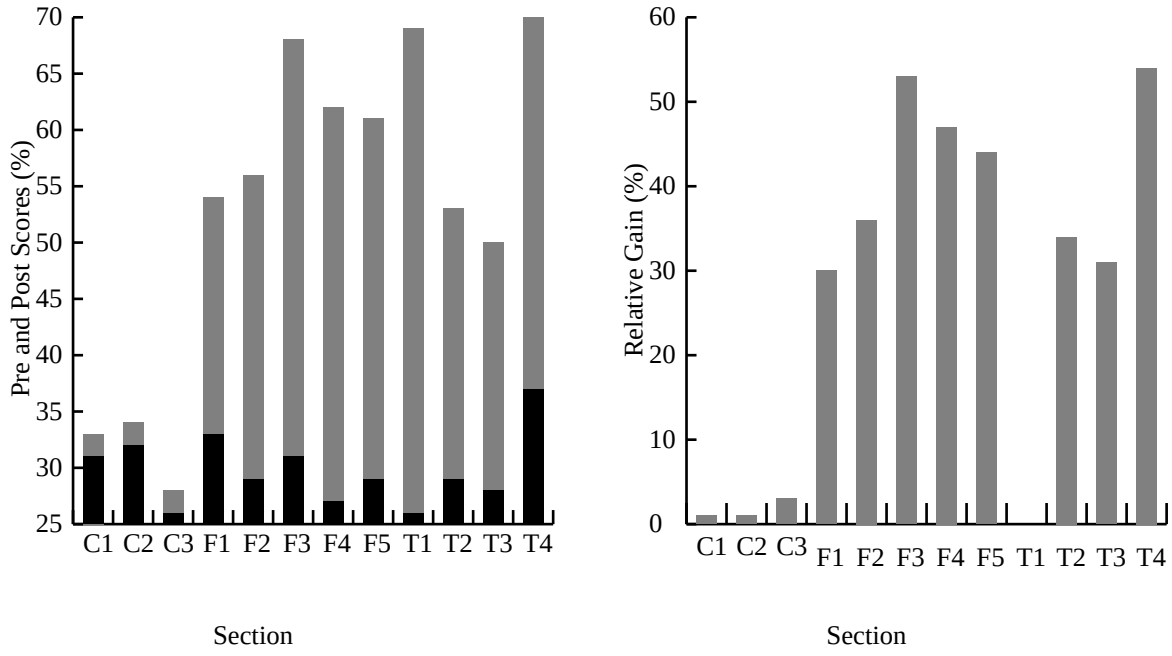


FIG. **1:** Survey results for the twelve sections utilized for this study. [left] Pre (dark) and post (light) scores; [right] relative gain.

strongly suggests that they utilized each other.[5] Thus, this section could not be used to make comparisons between sections.[6]

The second-biggest gain was obtained by section T4, which was the calculus-based section. As seen in figure 1, this population starts out with a higher score on the survey. Sections T2 and T3, on the other hand, was populated by the same type of students as those in the control and formative assessment sections. The impact of this bias is discussed in section b.

*b. Relative gain*

It can be difficult to compare pre-post differences among sections because the student population may differ among the sections.

To address this concern, a second parameter was examined. The second parameter, called the "normalized gain" (Hake, 1998), $g$, is obtained by dividing the gain (i.e., the difference between the pre and post scores) by the maximum possible gain (i.e., the difference between the pre score and the maximum possible score). This parameter is also known as the "effec-

[5] Almost all students had the same answers as at least one other student for all of the questions.

[6] In an attempt to compare section T1 with the other sections, students in sections T1 and F5 were identified in the second semester of the sequence the following semester (Spring 2007) and their mid-term grades were compared. Fourteen students from section F5 (with a GPA of 2.14) enrolled in the second semester and received an average midterm grade of 2.57, an increase of 0.43. In comparison, nineteen students from section T1 (with a GPA of 2.37) enrolled in the second semester and received an average midterm grade equal to the first semester average GPA (2.37). Although the sample size is small, this seems to suggest that the high post-test scores of section T1 are not valid.

tiveness index" (Hovland, Lumsdaine, and Sheffield, 1949) and the "gap closing parameter" (Ghery, 1972). Hake (1998) has shown that the normalized gain is a better measure of how effective an instructional methodology is because its value is highly uncorrelated with the pretest score. In other words, poor students should experience a higher change in score because the lower the initial score, the larger the maximum gain possible. Consequently, an increase (or decrease) of 10% has a bigger influence on the normalized gain if the pretest score is high (small maximum possible gain).

An additional challenge was to account for the change in student population between the pre and post offerings of the survey. As the semester progresses, some students drop the class or may not be present for the initial or final offerings of the survey. Furthermore, some students may inadvertently neglect to answer a question or two. To account for such changes, only those items answered on both the pre and post offerings were included in the computation of the normalized gain.

The normalized gain for for a particular section, then, was calculated by taking the average post score (using only items answered by students both pre and post) and subtracting the average pre score (also using only items answered by students both pre and post). This difference was then divided by the maximum possible can (100% minus the average pre score).

The average normalized gains for each section are shown in the right graph of Figure 1, with the results of section T1 removed.[7]

The signal observed previously still remains. The control sections (C1, C2 and C3) taught by the one-year temporary instructor, continue to show small relative gains ($g = +1\%$, $+1\%$ and $+3\%$) while the sections utilizing the formative assessment techniques (F1 through F5) show steady improvement in relative gains through the first three years ($g = +30\%$, $+36\%$, $+53\%$) with a small drop-off in subsequent years ($g = +47\%$, $+44\%$). Even with the drop-off, however, the relative gains are higher than those in the sections taught by the exemplary instructor (sections T2 and T3; $g = 34\%$ and $31\%$).

For comparison, in Hake's (1998) study using the Force Concept Inventory for the pre and post test, he found that the normalized gains of "traditional" courses tended to be around 23% ($\pm$ 4% std. dev.) whereas courses utilizing "interactive engagement" had normalized gains around 48% ($\pm$ 14% std. dev.), where he defines "interactive engagement" as "heads-on (always) and hands-on (usually) activities which yield immediate feedback through discussion with peers and/or instructors."

This is somewhat consistent with the findings of this study, as the traditional courses (C1-C3, T2-T3) experienced lower relative gains than those utilizing the formative assessment techniques (F3-F5). However, section T4, utilizing traditional techniques experienced a gain comparable to those of sections F3-F5. Thus, it isn't clear that the relative gain is independent of the pre-test score.

Possible reasons for the slight decline in relative gain from section F3 (53%) to section F5 (44%) are discussed in section 4.

---

[7] Not only was the post-test data invalidated by the the way it was administered (see discussion earlier), the pre-test was administered without having the students indicate their names. Without the names, it was not possible to match post-test answers with pre-test answers and so it was not possible to calculate relative gain.

## 4. Discussion

### *a. Textbook*

Recall that sections (T2 and T3) differed only by the textbook (Urone, 2001, vs. Cohen, 2004). The scores for both sections are similar, implying that the textbook has little effect.

On the other hand, upon closer examination of the results (not shown), it was found that for those areas stressed by the textbook, student scores improved whereas in areas not stressed by the textbook or poorly handled by the textbook, student scores decreased, leading to no change overall. The textbook is still under revision and may demonstrate a positive influence in future sections as its presentation in certain areas is strengthened. Indeed, within the five formative assessment sections (F1-F5), student evaluation of the textbook increased markedly from year to year. On the 5-point Likert scale used for student evaluations at ESU, student responses to the item "The textbooks contribute to my understanding of the subject" improved from 2.71 to 3.11 to 4.10 to 4.39 to 4.62, far exceeding the average for all faculty at ESU (3.95) or that obtained by the exemplary instructor in section T2 (3.75) where Urone (2001) was used.

It may also be the case that a different approach is necessary to take advantage of the textbook benefits (e.g., an approach that forces students to make better use of the textbook). Regardless, considering that the textbook lacked the color pictures, examples and ancillary materials common to traditional textbooks, it may be significant that it had no detrimental effect on student understanding.

### *b. Instructor Experience*

In every case, students taught by tenured instructors (F1-F5, T1-T4) outperformed those in the control sections (C1-C3), suggesting that tenured, continuing instructors may be more effective than temporary instructors.

Furthermore, the steady improvement in performance from section F1 to F3 suggests that instruction improves as the instructor gains proficiency in the formative assessment techniques. As the utilization of all three techniques were revised and refined each year, the scores on the evaluation instrument likewise improved, reaching the same level as those in sections T1 (where students were allowed to complete the survey at home) and T4 (where the students started out at a higher level).

It was also found that as the scores improved, so did the student evaluations of the instructor.[8] This suggests that variations in student evaluations may be a good indicator of student understanding *for a given instructor*. However, *across instructors*, the opposite was found. This can be seen by comparing sections T2 and F1 (taught by two different instructors during the fall of 2002). The performance of students was similar in both sections. Yet, the student evaluations for the instructor of section F1 were much worse than those for the instructor section T2.[9]

---

[8] For example, on the 5-point Likert scale used for student evaluations at ESU, student responses to the item "Overall, I rate this instructor a good teacher" went from 2.35 to 3.42 to 4.05 to 4.39 to 4.24 for sections T4 through T8.

[9] On the 5-point Likert scale used for student evaluations at ESU, student responses to the item "Overall, I

| Section | Semester | Class Participation | Email | HW | Relative gain |
|---|---|---|---|---|---|
| F1 | Fall-2002 | 64% | 16% | 39% | +30% |
| F2 | Fall-2003 | 63% | 36% | 54% | +36% |
| F3 | Fall-2004 | 72% | 41% | 63% | +53% |
| F4 | Fall-2005 | 76% | 1% | 60% | +47% |
| F5 | Fall-2006 | 70% | 42% | 63% | +44% |

**Table 3:** A summary of the participation rates for the five sections that used the formative assessment techniques.

The lower evaluation scores in sections F1 and F2 have been duplicated elsewhere, where it has been found that students in general become frustrated when they receive non-traditional instruction, even when such approaches result in improved conceptual understanding (see, e.g. Meltzer and Manivannan, 1996, and Crouch and Mazur, 2001).

And, in fact, the instructor made a conscious effort in sections F3 and F4 to address this concern. It isn't clear this purposeful action resulted in higher performance, but there was a strong correlation with the student evaluations.[10] In addition, during the last offering (section F5), when less of an emphasis was made to defend and explain the methodology, student evaluations declined, with a corresponding small decrease in student performance.

*c. Formative assessment techniques*

Another possible cause for the steady improvement in scores in sections F1 through F5 is the steady increase in the rate at which students participated in the formative assessment activities. Table 3 shows the average participation rates for the in-class remotes (class participation), email and homework. One can see a steady increase in student involvement that corresponds with a steady increase in relative gain.

By itself, the email activity appears to provide practically no benefit, since removing the email activity resulted in an insignificant decline (compare sections F3 and F4; email was still utilized for extra credit but few students took advantage of it). However, based upon the discussion above, it is possible that the scores would've continued to increase if the email activity was kept and so the lack of an increase from F3 to F4 may be more significant than one might surmise at first glance. As the instructor of those sections, the author sensed that the lack of email had a detrimental effect on student learning that was balanced by an improved textbook. However, further improvement was not found in the subsequent offering of section F5, even though email participation returned.

---

rate this instructor a good teacher" was 4.50 for the instructor of T2 whereas it was 2.35 for the instructor of F1.

[10] For example, on the 5-point Likert scale used for student evaluations at ESU, student responses to the item "The instuctor seems to care about my learning" went from 2.83 to 3.32 to 4.25 to 4.56 to 3.81 during the 5-year study.

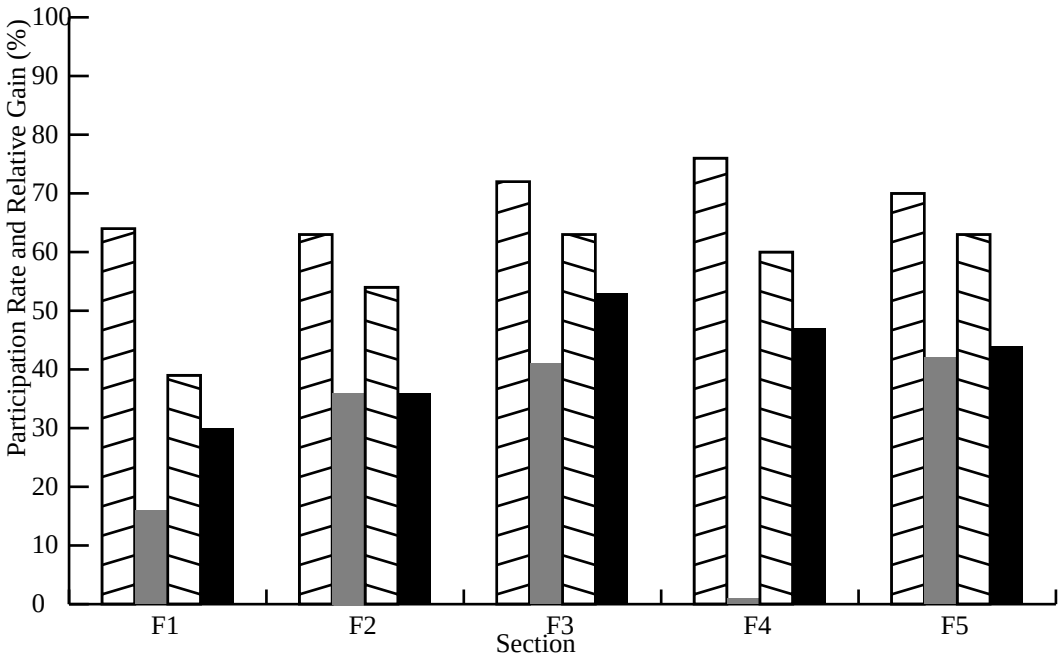


FIG. **2:** Participation rates for sections F1 through F5 in Class Participation (right-hatch), Email (light shading), homework (left-hatch) and relative gain (dark shading).

*d. Assessment Instrument*

Another aspect of the study that makes interpretation of the results a bit more difficult is that the assessment instrument was not rigorous. As mentioned in section 2, while many of the questions on the survey were selected from the Force Concept Inventory (Hestenes, Wells and Swackhamer, 1992; Halloun, Hake, Moscaand and Hestenes, 1995), several others were developed in-house. Due to the in-house development, the reliability and accuracy of the instrument is unknown. Analysis of the results, for example, revealed that two questions utilized misleading language. In one case, ambiguity in wording allowed for two answers to be interpreted as being correct. Since accepting both answers did not change the relative performances of each section, it is felt that the overall conclusions are still valid. However, in the other case, control sections C1-C3 outperformed all other sections. It is unclear whether the ambiguity in wording alone was responsible for the inverted scores.

## 5. Conclusions

Based upon this work, it seems we can make the following conclusions:

- Traditional physics instruction does not necessarily produce changes in student conceptual understanding (see, for example, sections C1-C3 in Figure 1).
- The instructor can have an effect on content understanding (compare section T2 with sections C1-C3 in Figure 1).

- Implementing formative assessment can lead to improved student understanding (see section F3 in Figure 1) beyond what might otherwise be possible.
- It may require several years to master the techniques (see, e.g., sections F1-F3), a time during which student evaluations may suffer even as student understanding improves.
- The textbook, by itself, does not affect student performance but may support other attempts to promote conceptual understanding.
- The email activity appears to provide an important piece of formative assessment but the other peices (email and textbook) together are still effective.
- The instructor needs to continuously and explicitly inform the students of the value of the formative assessment techniques used.

If this study could be continued, it would be desirable to examine whether the high performance of section F3 can be duplicated simply by consciously and continuously explaining the need for the techniques.

Unfortunately, the survey was distributed as a take-home exam in the fall of 2006 (see section T1) and this may have compromised its value as a way to make meaningful comparisons between sections. Thus, at the time, we don't foresee a way of continuing this content study past this year.[11]

[11] We may be able to identify whether the high performance of section T1 was dependent on allowing students to take home the survey. This would involve the creation of a new survey. Since the performance of students in the sections utilizing the formative assessment techniques (see sections F3 through F5) is somewhat uniform, we could probably assume that performance will be somewhat the same in a future offering of such a section. We can then use that as a benchmark against other sections.

## REFERENCES


Cohen, R. A., 2002: *The Fundamentals of College Physics*, Vol. I, Version 3.0, East Stroudsburg University.

———-, 2003: *The Fundamentals of College Physics*, Vol. I, Version 4.0, East Stroudsburg University.

———-, 2004: *The Fundamentals of College Physics*, Vol. I, Version 5.0, East Stroudsburg University.

———-, 2005: *The Fundamentals of College Physics*, Vol. I, Version 6.0, East Stroudsburg University.

———-, 2006: *The Fundamentals of College Physics*, Vol. I, Version 7.0, East Stroudsburg University.

Crouch, C. H. and E. Mazur, 2001: Peer Instruction: Ten years of experience and results, *Am. J. Phys.*, **69**, 970–977.

Cutnell, J. D., and K. W. Johnson, 2004: *Physics*, sixth edition, Hoboken, NJ: John Wiley & Sons, Inc.

Ghery, F. W. 1972: Does mathematics matter, pp. 142–157 in *Research papers in economic education*, A. Welch, editor, Joint Council on Economic Education, New York, New York, USA.

Hake, R. R., 1998: Interactive-engagement vs traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses, *Am. J. Phys.*, **66**(1), 64–74.

Halloun, I., R. R. Hake, E. P. Mosca and D. Hestenes, 1995: *Force Concept Inventory* (Revised 1995); online (password protected) at http://modeling.la.asu.edu/R&E/Research.html.

Hestenes, D., M. Wells, and G. Swackhamer, 1992: Force Concept Inventory, *Phys. Teach.*, **30**, 141–158.

Hovland, C. I., A. A. Lumsdaine and F. D. Sheffield, 1949: *A baseline for measurement of percentage change*, in *Experiments on mass communication*, C. I. Hovland, A. A. Lumsdaine, and F. D. Sheffield, editors, 1965, Wiley (first published in 1949), reprinted as pages 77–82 in *The language of social research: a reader in the methodology of social research*, P. F. Lazarsfeld and M. Rosenberg, editors, 1955, Free Press, New York, New York, USA.

Knight, R. D., 2004: Physics for Scientists and Engineers: a strategic approach, San Francisco, CA: Pearson Education, Inc.

Lawson, A. E., 1978: Development and validation of the classroom test of formal reasoning, *J. Res. Sci. Teach.*, **15**(1), 11–24.

Meltzer, D. E. and K. Manivannan. 1996: Promoting Interactivity in Physics Lecture Classes, *Phys. Teach.*, **34**, 72–76.

Novak, G. M., E. T. Patterson, A. D. Gavrin and W. Christian, 1999: *Just-in-Time Teaching*, Upper Saddle River, NJ: Printice-Hall, Inc.

Urone, P. P., 2001: *College Physics*, second edition, Pacific Grove, CA: Brooks/Cole.