CHAPTER 5

Evaluating Our Instruction: Surveys

Mathematics may be compared to a mill of exquisite
workmanship, which grinds you stuff of any degree of fineness;
but, nevertheless, what you get out depends on what you put in;
and as the grandest mill in the world
will not extract wheat flour from peascod,
so pages of formulae will not get a definite result
out of loose data.
T. H. Huxley [Huxley 1869]

As I discussed in the last chapter, there are two ways to probe what is happening in one's class. One way is to assess how much each student has learned in order to decide the extent to which that student receives a public certification of his or her knowledge—a grade. A second way is to probe our class overall in order to determine whether the instruction we are delivering is meeting our goals. I refer to the first as *assessment*, and the second as *evaluation*. In the last chapter we discussed how to assess each student so as to see what he or she was learning. In this chapter, we discuss how to get a snapshot of how the class is doing overall for the purpose of evaluating our instruction.

As a result of the context dependence of the cognitive response (Principle 2 in chapter 2), in some contexts students may choose to use the model they are being taught, while in other contexts they may revert to using more naïve resources. When we look at a class broadly (especially a large class), we can tolerate larger fluctuations in individual responses than we can when we are assessing individual students. The students' individualized choices of what ideas to use and when to use them depend on uncontrollable and unknowable variables (their internal mental states). As a result, their answers may appear random on a small set of closed-end questions on each topic to be probed. But these same few questions may give a good average view of what is happening, despite not giving a complete picture of the knowledge of any single student.

RESEARCH-BASED SURVEYS

A cost-effective way to determine the approximate state of a class's knowledge is to use a carefully designed research-based survey. By a *survey* I mean a reasonably short (10- to 30-minute) machine-gradable test. It could consist of multiple-choice or short-answer questions, or it could contain statements students are asked to agree or disagree with. It can be delivered on paper with Scantron™ sheets or on computers.[1] It can be delivered to large numbers of students and the results manipulated on computers using spreadsheets or more sophisticated statistical analysis tools.

By *research-based*, I mean that the survey has been developed from qualitative research on student difficulties with particular topics and has been refined, tested, and validated by detailed observations with many students. Broadly, to achieve good surveys (surveys that are both valid and reliable—see the discussion below) requires the following steps.

- Conduct qualitative research to identify the student models underlying their responses.
- Develop a theoretical framework to model the student responses for that particular topic.
- Develop multiple-choice items to elicit the range of expected possible answers.
- Use the results—including the student selection of wrong answers—to facilitate the design of new instructions as well as new diagnostic and evaluation tools.
- Use the results to guide construction of new qualitative research to further improve the survey.

This process places development of evaluational tools firmly in the research-redevelopment cycle of curriculum construction and reform discussed in more detail in chapter 6 (Figure 6.1).

Surveys may focus on a variety of aspects of what students are expected to learn in both the explicit and hidden curriculum. *Content surveys* probe student knowledge of the conceptual bases of particular content areas of physics. *Attitude surveys* probe student thinking about the process and character of learning physics. Over the past two decades, physics education researchers have developed dozens of surveys that probe topics from mechanics (the Force Concept Inventory) to the atomic model of matter (the Small Particle Model Assessment Test). Seventeen such surveys are included on the Resource CD accompanying this volume. They are listed in the Appendix at the end of this volume.

Why use a research-based survey?

Sagredo scoffs at my emphasis on creating a survey through research. "I've given machine-gradable multiple-choice final exams in my large classes for years. Don't my grades count as course evaluations?" Certainly they do, Sagredo. But there are dangers in interpreting exam results as course evaluations.

The questions we choose often are constrained by a number of factors that may be unrelated to student learning. The first danger is that there is pressure from students (and sometimes from administrations) to have an "appropriate" grade distribution. Students are

[1] Studies to look for differences between paper-delivered and computer-delivered surveys have so far had ambiguous results.

comfortable with class averages near 80%, with 90% being the boundary for an A and with few grades falling below 60%. Although this may make students and administrations happy, it presses us to produce the requisite number of As no matter what our students have learned.[2]

A second danger arises because we are interested in what our students have really learned, not in what they think you want them to say. By the time they get to university, many students have become quite adept at "test-taking skills." These are even taught in some schools in order to help students (and school administrators) receive higher evaluations. I'm not criticizing students for taking this approach. I made use of them myself when I took standardized tests. Students taking an exam have the goal to obtain the highest possible score given what they know. Instructors want the score to accurately reflect what their students know. If we are not aware in detail of our students' starting states—what resources and facets are easily activated—we might be hard pressed to come up with reasonable wrong answers for a multiple-choice test or with tempting and misleading cues for short-answer questions. Without these *attractive distractors,* students can focus on eliminating obviously wrong answers and can use their test-taking skills to get a correct result even if they only have a very weak understanding of the subject.

Neither of these dangers is trivial. The other side of the first danger is that instruction is never purely objective. It is oriented, in principle, to achieving goals set by the instructor, though sometimes those goals are tacit, inconsistent, or inappropriate to the particular student population involved. An instructor's exams should reflect his or her own particular learning goals for his or her students. What is appropriate for us to demand our students learn is a continuing negotiation among instructors, their students, and outside pressures such as administrators, parents, and faculty in the departments our courses serve.

If an instructor is unaware of common student confusions or of how students tend to respond to particular questions, the result on questions she creates for an exam may not reflect what she thinks it does. Furthermore, without a carefully developed question based on research and a clear understanding of common naïve responses, the students' wrong answers may provide little useful information beyond "my students don't know the answer."

In trying to interpret the responses of students on closed exam questions, we may encounter one or more of the following problems.

1. If a multiple-choice test does not have appropriate distractors, you may not learn what the students really think.
2. The fact that students give the right answer does not mean they understand why the answer is right.
3. Since student responses are context dependent, what they say on a single question only tells part of the story.
4. Problems in the ordering or detailed presentation of the distractors may cause problems in interpreting the results.

[2] In my classes, I try to set exams that have averages between 60% and 65%. A grade over 75% is considered an A. At this level of difficulty, even the good students get some feedback about where they need to improve. See my model of examination delivery discussed in chapter 4.

5. It's easy to overinterpret the implications of a single relatively narrow test—especially if it only has one style of question.

Concept surveys that are carefully constructed with these points in mind can provide a useful tool as part of our evaluation of our instruction.

Surveys and the goals of a class

While giving a lecture on physics education research to colleagues in a neighboring physics department, I once showed some of the questions from the Force Concept Inventory (discussed in detail below and given on the Resource CD). One faculty member objected quite vigorously. "These are trick questions," he said. "What do you mean by a 'trick question'?" I asked. He answered, "You really have to have a deep understanding of the underlying physics to answer them correctly." After a substantial pause, allowing both him and the rest of the audience to consider what had just been said, I responded, "Exactly. I want all the questions I ask to be trick questions."

This raises a deep question. What is it we want our students to learn from our instruction? My colleague clearly had much lower expectations for his students than I did—in one sense. He was satisfied with recognition of an answer but didn't care if his students could not distinguish between the correct (physics) answer and an attractive (but incorrect) commonsense alternative. On the other hand, he probably demands much more in the way of sophisticated mathematical manipulations on examinations than I do and is satisfied if his students can match a complex problem-solving pattern to one they have memorized. I do not care if my students can pattern match complex mathematical manipulations. I want them to be able both to formulate physics problems out of real-world situations and to interpret their answers sensibly. If we could attain both goals in a one-year course, I would be delighted, but at present, I don't know how to do it given the time and resource constraints of a large class.

This shift in goals can produce some difficulties. Sagredo and I both teach algebra-based physics on occasion. When he looks at my exams, he complains that they are too easy and that I'm "dumbing-down" the course. Interestingly enough, many students have reported to me that the scuttlebutt among the students is "take Sagredo if you want an easy A" and that my course is the one to take "if you want to work hard and really understand it." Whenever Sagredo agrees to give one of my questions to his students on an exam, he is surprised at how poorly they do. My students would also do poorly on some of his questions.

In the end, when the chalk meets the blackboard, each individual instructor defines his or her own goals. Nonetheless, there is clearly a need for a community to form to discuss both the appropriate goals for physics instruction and how to evaluate the extent to which those goals are reached. That is why I favor the use of research-based surveys as one element in our evaluations of our instructional success. They are explicit in what they are evaluating, they are based on careful study of student difficulties, and they are carefully tested for validity and reliability.

Delivering a survey in your class

Whenever possible, I give pre-post surveys (i.e., at the beginning and end of the class). In some classes that have an associated lab, the first and last weeks of class do not have labs, and so I tell students to come in then to take surveys. In classes that do not have a lab or a blank week, I am willing to take time in the first and last classes to give surveys. To encourage students to take them (and to make them up if they miss the class), I give everyone who completes each survey 5 grade points (out of a total of about 1000 points for the class as a whole). If everyone does them, the surveys have no impact on the grading pattern.

To analyze a survey, it is important to compare only *matched data sets*. That is, only students who take both the pre- and the post-tests should be included. This is because there may be biases in the populations who take the two tests. For example, students who drop the course in the middle would take the pre-test and not the post-test. If the group of students dropping the class were biased toward the lower scoring students on the pre-test, this would bias the pre-post comparison toward high gains. At least if a matched set is used, one is looking at the true gains for a particular set of students.

The danger discussed in the previous section—that students often give us what they think we want instead of what they think—has three important implications for how surveys should be delivered, especially if the survey is to meet the purpose of evaluating our instruction rather than certifying the students. The three implications are:

- We have to be careful to "teach the physics" but not "teach to the test."
- Survey solutions should not be distributed or "gone over" with the students.
- Surveys should be required (given credit) but not graded.

The first implication, not teaching to the test, is a delicate one. We want the test to probe students' knowledge appropriately, and we want our instruction to help them gain the knowledge that will be probed. Why then is "teaching to the test" usually considered such a pejorative? I think that it is because in this case we are implicitly using a fairly sophisticated model of student learning: students should learn how to think, not to parrot back answers they don't understand. In our cognitive model (described in chapter 2), this idea can be expressed more explicitly by saying that students should develop a strong mental model of the physics with many strong associations that will permit them to identify and use appropriate solution techniques to solve a wide variety of problems presented in diverse contexts. Research strongly demonstrates that when students learn an answer to a problem narrowly, through pattern matching, small changes in the problem's statement can lead to their being unable to recognize the pattern.[3] So if during instruction we give students the specific question that will appear on the test, framed exactly as it will be framed there, I call it *teaching to the test.*

This leads to the second implication: Survey solutions should not be posted or given out to students. Research-based survey items can take a long time to develop. Students have to be interviewed to see how they are reading and interpreting the items and their answers. Surveys have to be delivered repeatedly to study distributions and reliability at a cost to class time. A carefully developed survey, whatever limitations it may have, is an extremely valuable resource for the community of physics teachers, but it is fragile. If they are graded and the answers are posted or discussed in class, they spread—to fraternity/sorority solution

[3] I have seen students who solved problems by pattern matching fail to recognize a problem they knew if the picture specifying the problem was reversed (mirror image).

collections and to student websites—and become a property of the student community rather than of the instructional community. They are transformed from a moderately effective evaluation tool for the teacher to a "test-taking skills" tool for the student.

This leads directly to the third implication: Surveys should not be a part of the student's grade. This is a somewhat controversial point. Sagredo suggests that students will not take a test seriously unless it is graded. This may be true in some populations. At Maryland, it has been my experience that 95% of my students hand in surveys that show they have been thought through and answered honestly.[4] This might differ in other populations. If a test is graded, at least some students will make a serious effort to find out the correct answers and perhaps spread them around. Since I very much don't want my students to do this, I treat my exams (which I consider pedagogical tools to help facilitate student learning) and my surveys (which I consider evaluational tools to help me understand my instruction) differently.

There is an additional reason for leaving surveys ungraded. Students often use their test-taking skills to try to produce an answer that they think the teacher will like, even if they don't really think that is the answer. A graded exam definitely tends to cue such responses. I am more interested in finding out how students respond when such motivation is removed in order to see whether instruction has had a broader impact on student thinking.

For a survey to be useful, it should be both valid and reliable. I discuss these conditions next. In the remainder of the chapter I discuss two kinds of surveys that have been developed and that are currently widely available: content surveys and attitude surveys.

UNDERSTANDING WHAT A SURVEY MEASURES: VALIDITY AND RELIABILITY

In order to be a useful tool in evaluating instruction, a survey should be *valid*; that is, it should measure what it claims to measure. A survey should also be *reliable;* that is, it should give reproducible results. When we're talking about measurements of how people think about physics instead of about measurements of physical properties, we have to consider carefully what we mean by these terms.

Validity

Understanding the validity of a survey item, either in a content or attitude survey, is not as trivial as it may appear on the surface. What's in question is not just the issue of whether the physics is right, but whether the question adequately probes the relevant mental structures. To see what this means, consider the following example. The most common student confusion about velocity graphs is whether they should "look like the motion" or like the rates of change of the motion. If you ask students in introductory university physics "which graph corresponds to an object moving away from the origin at a uniform rate" (as in the problem shown in Figure 5.4) and provide them with the correct (constant) graph but <u>not</u> with the choice of the attractive distractor (the linearly increasing graph), you will get a high score but an invalid question. This one is especially subtle. Jeff Saul found that if both of these graphs

[4] Evidence to the contrary might be: all answers the same, answers in a recurring pattern, survey completed in one-fourth the average time, and so on.

were included but the correct answer given first, ~80% of the students selected the right answer. But if the attractive distractor (a rising straight line) was given first, the success rate fell by about a factor of 2 [Saul 1996].

In order to achieve validity, we need to understand quite a bit not only about the physics but also about how students think about the physics. Since human beings show great flexibility in their responses to situations, we might expect an intractably large range of responses. Fortunately, in most cases studied, if the topic is reasonably narrowly defined, a fairly small number of distinct responses (two to ten) accounts for a large fraction of the answers and reasonings that are found in a given population. Understanding this range of plausible variation is absolutely essential in creating valid survey items. As a result, the first step in creating a survey is to study the literature on student thinking and do extensive qualitative research to learn how students think about the subject.

Even if we understand the range of student thinking on a topic, for an item to be valid students must respond to it in the expected way. This is one of the most frustrating steps in creating a survey. A good way to probe this is to observe a large number of students "thinking aloud" while doing the survey. Culture shifts and vocabulary differences between the (usually middle-aged) test designers and (usually young adult) subjects can produce dramatic surprises. In our validation interviews for the Maryland Physics Expectations (MPEX) survey discussed later in this chapter, we wanted to know if students understood our motivation for doing derivations in lecture. To our dismay we learned that a significant fraction of our calculus-based introductory physics students were unfamiliar with the word "derivation" and thought it meant "derivative."[5] To get a consistent valid response we had to rephrase our items.

Reliability

Reliability also has to be considered carefully. In probing human behavior, this term replaces the more standard physics term *repeatability.* We usually say that repeatibility means that if some other scientists repeated our experiment, they would get the same result, within expected statistical variations. What we really mean is that if we prepare a new experiment in the same way as the first experimenter did, using equivalent (but not the same) materials, we would get the same result. We don't expect to be able to measure the deformability of a piece of metal many times using the same piece of metal. We are comfortable with the idea that "all muons are identical," so that repeating a measurement of muon decay rates doesn't mean reassembling the previously measured muons out of their component parts.

But when it comes to people, we are accustomed to the idea that people are individuals and are not equivalent. If we try to repeat a survey with a given student a few days later, we are unlikely to get the identical result. First, the "state of the student" has changed somewhat as a result of taking the first survey. Second, the context dependence of the cognitive response reminds us that "the context" includes the entire state of the student's mind—something over which we have little control. Experiences between the two tests and local situations (Did a bad exam in another class make her disgruntled about science in general? Did an argument with

[5] This is perhaps a result of the unfortunate strong shift away from "proof" in high school math classes that took place in the 1990s.

his girl friend last night shift his concerns and associations?) may affect student responses. And sometimes, students do, in fact, learn something from thinking about and doing a survey.

Fortunately, these kinds of fluctuations in individual responses tend to average out over a large enough class. Measures can become repeatable (reliable) when considered as a measure of a population rather than as a measure of an individual. However, we must keep in mind that according to the individuality principle (Principle 4, chapter 2), we can expect a population to contain substantial spreads on a variety of measures. Any survey is measuring a small slice of these variables. There is likely to be a significant spread in results, and the spread is an important part of the data.[6]

In the educational world, reliability testing is sometimes interpreted to mean that students should respond similarly to the same question formulated in different ways. For example, one might present an "agree–disagree" item in both positive and negative senses. For the MPEX, discussed below, we have the following pair of items to decide whether a student feels that she needs to use her ordinary-world experiences in her physics class.

> *To understand physics, I sometimes think about my personal experiences and relate them to the topic being analyzed.*
>
> *Physics is related to the real world and it sometimes helps to think about the connection, but it is rarely essential for what I have to do in this course.*

Although these items are designed to be the reverse of each other, they are not identical. The difference between "sometimes" and "rarely essential" can lead a student to split the difference and agree with both items, especially if that student is on the fence or is in transition. Even when items are closer than these, students can hold contradictory views. In this case, the "lack of reliability" in the responses to the matched questions lies not in the test but in the student. Care must be taken not to eliminate questions that show this kind of "unreliability," lest one bias the survey toward only seeing topics on which most students have formed coherent mental models.

CONTENT SURVEYS

In the remainder of this section I discuss three of the most commonly used surveys in mechanics in detail: the Force Concept Inventory (FCI), the Force and Motion Conceptual Evaluation (FMCE), and the Mechanics Baseline Test (MBT). The additional content surveys that are included on the CD provided with this book are listed and described briefly in the Appendix.

The FCI

One of the most carefully researched and most extensively used concept surveys in our current collection is the Force Concept Inventory (FCI) developed by David Hestenes and his collaborators at Arizona State University [Hestenes 1992a]. This is a 30-item multiple-choice

[6] A useful metaphor for me is a spectral line. For many physical circumstances, the width and shape of the line is important data, not just its centroid.

survey meant to probe student conceptual learning in Newtonian dynamics. It focuses on issues of force (though there are a few kinematics questions), and it is easily deliverable. Students typically take 15 to 30 minutes to complete it.

Building a probe of student conceptual understanding requires both understanding the fundamental issues underlying the physics to be probed, as viewed by the professional scientist, and understanding the common naïve conceptions and confusions the students spontaneously bring to the subject as a result of their experience. In creating their mechanics concept test, Hestenes and his collaborators first thought carefully about the conceptual structure of Newtonian mechanics. But understanding the professional's view is not enough. The test has to be designed to respond properly when considered from the *student's* point of view. The distractors (wrong answers) should distract! That is, there should be answers that correspond to what many naïve students would say if the question were open ended and no answers were given.

Hestenes and his collaborators relied on existing research and did extensive research of their own to determine spontaneous student responses [Halloun 1985a] [Halloun 1985b] [Hestenes 1992a]. They then compiled a list of common naïve conceptions and attempted to create questions that would reveal whether or not the students harbored these naïve conceptions. Their list of naïve conceptions is given in the FCI paper [Hestenes 1992a]. Many of them are directly related to the facets created applying primitive reasoning in the context of motion. (See chapter 2.)

Finally, in constructing the FCI, Hestenes and his collaborators chose to use semirealistic situations and everyday speech rather than technical physics speech in order to set the context to be the student's personal resources for how the world works rather than what one is supposed to say in a physics class. See, for example, the upper part of Figure 2.6, and Figure 5.1. Answer (C) in Figure 5.1 is an example of a research-based distractor. Few physics instructors who have not studied the research literature would think of choosing such an item;

Imagine a head-on collision between a large truck and a small compact car. During the collision:

(A) the truck exerts a greater amount of force on the car than the car exerts on the truck.
(B) the car exerts a greater amount of force on the truck than the truck exerts on the car.
(C) neither exerts a force on the other; the car gets smashed simply because it gets in the way of the truck.
(D) the truck exerts a force on the car but the car does not exert a force on the truck.
(E) the truck exerts the same amount of force on the car as the car exerts on the truck.

Figure 5.1 A question from the Force Concept Inventory [Hestenes 1992a].

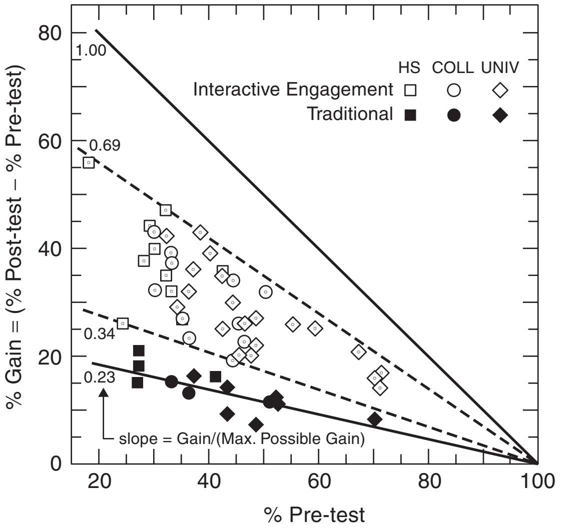


Figure 5.2 A plot of class average pre-test and post-test FCI scores for a collection of classes in high school, college, and university physics classes using a variety of instructional methods [Hake 1992].

it's not even "on their screen" as a possible wrong answer. But a significant number of naïve students, unaccustomed to seeking forces from objects as the cause of all changes in motion, actually select this answer.[7]

The FCI is perhaps the most widely used concept survey in the nation today. Its publication in 1992 stirred a great deal of interest in physics education reform among the community of college and university physics instructors. Looking at the test, most faculty declared it "trivial" and were shocked when their students performed poorly.[8] A typical score for a class of entering calculus-based physics students is 40% to 50% and a typical score for a class of entering algebra-based physics students is 30% to 45%. At the completion of one semester of mechanics, average scores tend to rise to about 60% for calculus-based students and 50% for algebra-based students. These are rather modest and disappointing gains.

Richard Hake of Indiana University put out a call for anyone who had given the FCI pre-post in a college class to send him their results, together with a description of their class. He collected results from over 60 classes [Hake 1992]. His results are displayed in Figure 5.2 and show an interesting uniformity. When the class's gain on the FCI (post-test average–pre-test average) is plotted against the class's pre-test score, classes of similar structure lie approximately along a straight line passing through the point (100,0). Traditional classes lie on the line closest to the horizontal axis and show limited improvement. The region between

[7] College students who have previously taken high school physics are less likely to choose (C) as an alternative here. They are more likely to select an "active agent" primitive or a facet built on a "more is more" primitive and to select (A).

[8] Compare the Mazur story in chapter 1. Mazur was influenced by the earlier version of the Hestenes test [Halloun 1985a], as was I.

the two dotted lines represents classes with more self-reported "active engagement." Hake claims that classes lying near the line falling most steeply reported that they were using active-engagement environments and a research-based text. This suggests that the negative slope of the line from a data point to the point (100,0) is a useful figure of merit:

$$g = (\text{class post-test average} - \text{class pre-test average})/(100 - \text{class pre-test average})$$

where the class averages are given in percents.

The interpretation of this is that two classes having the same figure of merit, g, have achieved the same *fraction of the possible gain*—a kind of educational efficiency. Hake's results suggest that this figure of merit is a way of comparing the instructional success of classes with differently prepared populations—say, a class at a highly selective university with entering scores of 75% and a class at an open enrollment college with entering scores of 30%. This conjecture has been widely accepted by the physics education research community.[9]

Hake's approach, though valuable as a first look, leaves some questions unanswered. Did people fairly and accurately represent the character of their own classes? Did a selection occur because people with good results submitted their data while people with poor results chose not to? To answer some of these questions, Jeff Saul and I undertook an investigation of 35 classes at seven different colleges and universities [Redish 1997] [Saul 1997]. Four different curricula were being used: traditional, two modest active engagement reforms (Tutorials and Group Problem Solving: ~one hour of reform class per week), and a deeply reformed high active-engagement curriculum (Workshop Physics) in early implementations.[10] We gave pre-post FCI in each class and observed the classes directly. The FCI results are summarized in Figure 5.3.

These results confirm Hake's observations and give support to the idea that g is one plausible measure of overall gain. Some additional interesting conjectures may be made after studying this figure.

1. In the traditional (low-interaction lecture-based) environment, what the lecturer does can have a big impact on the class's conceptual gains.

The peak corresponding to the traditional class is very broad. At Maryland, where we observed the largest number of classes in a reasonably uniform environment, the classes with the largest gains were taught by award-winning professors who tried to actively engage their classes during lecture. The lowest gains were taught by professors who had little interest in qualitative or conceptual learning and focused their attention on complex problem solving. (For the detailed "unsmoothed" version of the Maryland results, see Figure 8.3.)

2. In the moderate active-engagement classes (one modified small-class group-learning hour per week), much of the conceptual learning relevant to FCI gains was occurring in the modified class.

[9] Some detailed preliminary probes of this question have been reported at meetings of the American Association of Physics Teachers (AAPT), but no decisive publications have yet resulted.

[10] See chapters 8 and 9 for detailed descriptions of the methods.

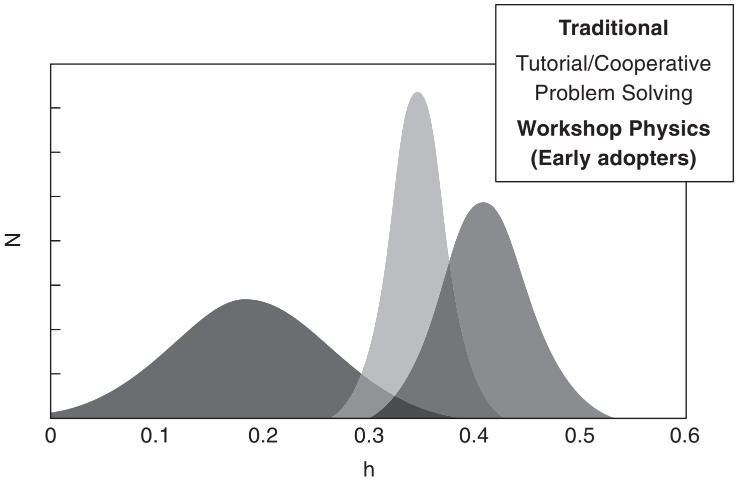


Figure 5.3 A plot of the fractional FCI gain achieved in three types of classes: traditional, moderate active engagement (tutorial/group problem solving), and strong active engagement (early adopters of workshop physics). Histograms are constructed for each group and fit with a Gaussian, which is then normalized [Saul 1997].

This is suggested by the narrowness of the peak and the fact that it lies above the results attained by even the best of the instructors in the traditional environments.

3. Full active-engagement classes can produce substantially better FCI gains, even in early implementations.

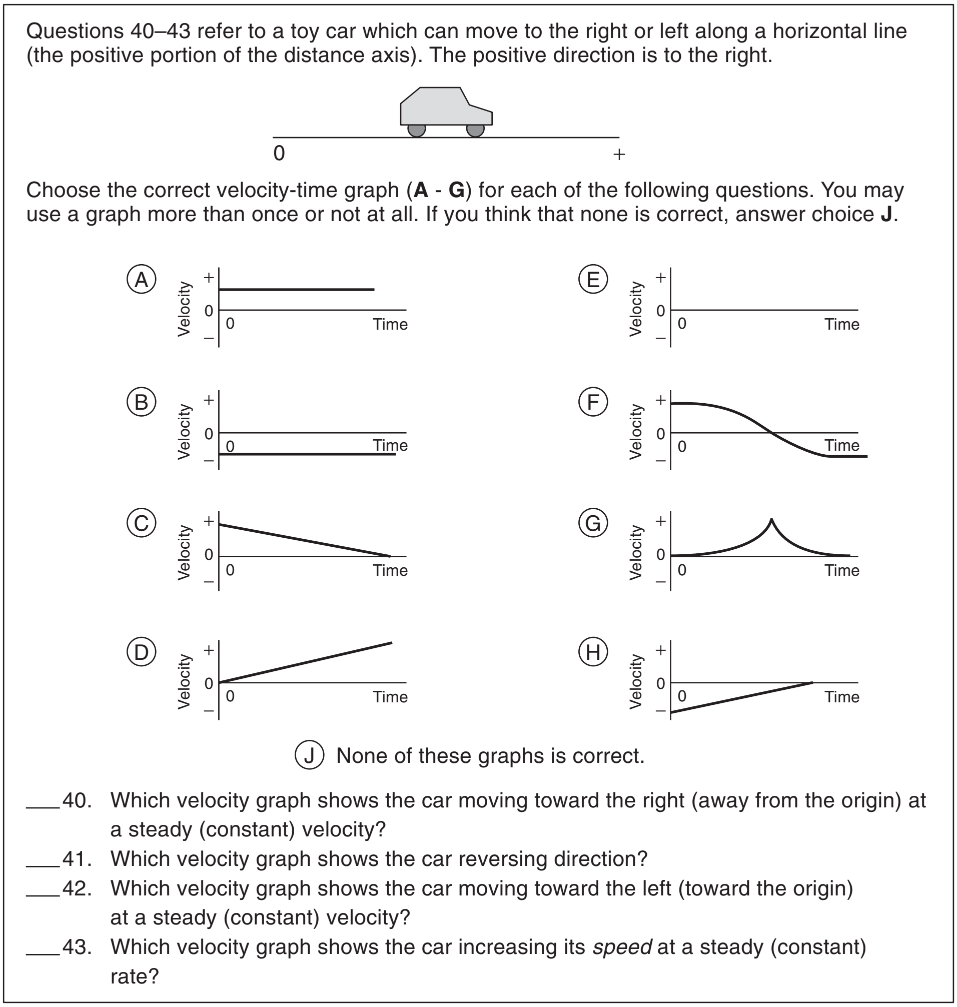
This is suggested by the results from the Workshop Physics classes studied. For a more detailed discussion of this issue (and for the results from mature Workshop Physics at the primary site), see chapter 9.

Although the FCI has been of great value in “raising the consciousness” of the community of physics teachers to issues of student learning, it has its limitations. Besides those limitations associated with all machine-gradable instruments, it is lacking in depth on kinematics issues that are to some extent a prerequisite to understanding the issues probed. A more comprehensive survey is provided by the Force and Motion Conceptual Evaluation.

The FMCE

The Force and Motion Conceptual Evaluation (FMCE) was developed by Ron Thornton and David Sokoloff [Thornton 1998]. In addition to the dynamical issues stressed by the FCI, this survey addresses student difficulties with kinematics, especially difficulties with representation translation between words and graphs. It is longer than the FCI, with 47 items in a multiple-choice multiple-response format that is somewhat more difficult for students to untangle than the (mostly) straightforward FCI multiple-choice items. As a result, students need more time to complete the FMCE—from 30 minutes to an hour.

An example of an FMCE item is given in Figure 5.4. Although this looks superficially trivial, students have a great deal of difficulty in choosing the correct graphs until they have



Questions 40–43 refer to a toy car which can move to the right or left along a horizontal line (the positive portion of the distance axis). The positive direction is to the right.

0
+

Choose the correct velocity-time graph (**A** - **G**) for each of the following questions. You may use a graph more than once or not at all. If you think that none is correct, answer choice **J**.

A
B
C
D
E
F
G
H
Velocity
+
0
–
0
Time

Ⓙ None of these graphs is correct.

___40. Which velocity graph shows the car moving toward the right (away from the origin) at a steady (constant) velocity?

___41. Which velocity graph shows the car reversing direction?

___42. Which velocity graph shows the car moving toward the left (toward the origin) at a steady (constant) velocity?

___43. Which velocity graph shows the car increasing its *speed* at a steady (constant) rate?

Figure 5.4 An MCMR set of items from the FMCE [Thornton 1998].

clearly differentiated the concepts of velocity and acceleration and have developed good graph-mapping skills.[11] (In my own use of this survey, I exchange graphs (A) and (D) so as to better probe how many students are tempted to assign the "linearly rising" graph to a constant velocity.)

The FMCE is structured into clusters of questions associated with a particular situation, as shown in Figure 5.4. This tends to "lock" students into a particular mode of thinking for the cluster of problems and may not give a clear picture of the range of student confusion on a particular topic [Bao 1999].

[11] By "graph-mapping" skills, I mean the ability to map a physical situation onto a variety of different graphical representations.

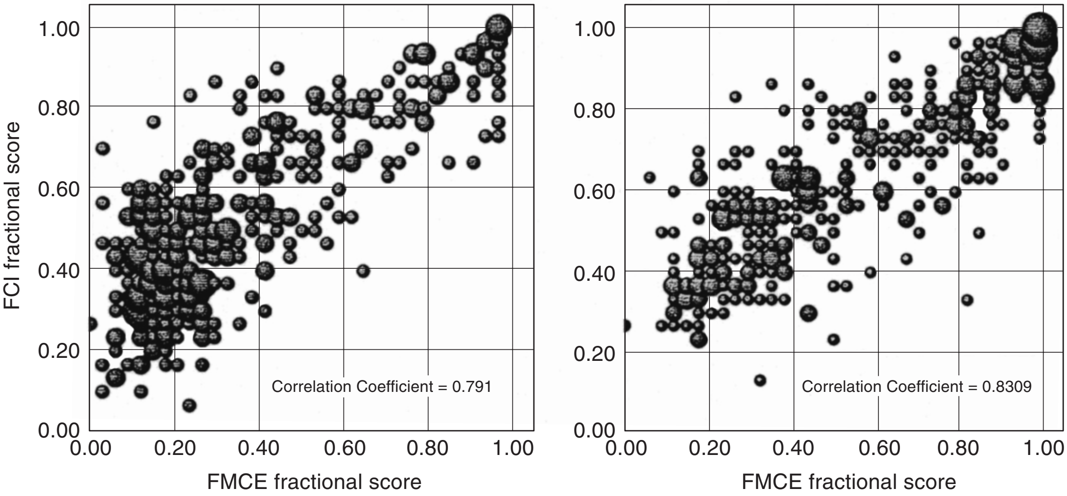


Figure 5.5 Scatter plot of FMCE versus FCI scores pre (left) and post (right). The size of the markers indicates the number of students with those scores [Thornton 2003].

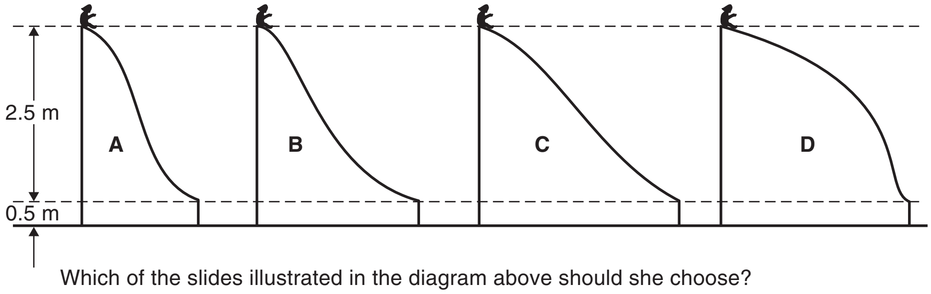
Some of the items in the FMCE are included to "set up" the student's frame of mind or to check for students who are not taking the test seriously (e.g., item 33). All students are expected to get these correct since they cue widely held facets that lead to the correct result. See the description of the FMCE analysis in the file on the CD. (The FMCE analysis template included on the CD is already set up to handle this.)

Thornton and his collaborators have carried out extensive studies of the correlation between FMCE results and FCI results [Thornton 2003]. They find a very strong correlation ($R = 0.8$) between the results, but the FMCE appears to be more challenging to low-scoring students, with few students scoring below 25% in the FCI while FMCE scores go down to almost 0%. Scatter plots of pre- and post-FCI versus FMCE scores are shown in Figure 5.5. (The areas of the circles are proportional to the number of students at the point.)

The MBT

Both the FCI and the FMCE focus on components of basic conceptual understanding and representation translation. Our goals for physics classes at the college level usually include applying conceptual ideas to solve problems. Hestenes and his collaborators created the Mechanics Baseline Test (MBT) to try to probe students' skill at making these connections [Hestenes 1992b]. Scores tend to be lower than on the FCI. Although this survey is designed for the introductory physics class, David Hestenes told me that when he gave it to the physics graduate students in his first-year graduate classical mechanics class, it correlated well with their grades. An example of an MBT item is given in Figure 5.6. In this item, students have to recognize the relevance of energy conservation. Students who fail to do so tend to activate various facets or other associations.

10. A young girl wishes to select one of the **frictionless** playground slides illustrated below to give her the greatest possible speed when she reaches the bottom of the slide.



Which of the slides illustrated in the diagram above should she choose?

(A) A (B) B (C) C (D) D
(E) It doesn't matter; her speed would be the same for each.

Figure 5.6 An item from the MBT [Hestenes 1992b].

ATTITUDE SURVEYS

If we want to understand whether our students are making any progress on our hidden curriculum of learning both process and scientific thinking, we need to find some way to probe the state of their attitudes.[12] One approach that has provided a useful first look is to use an attitude survey. Three attitude surveys are provided on the CD accompanying this book: the MPEX, the VASS, and the EBAPS.

In using an attitude survey, one needs to be aware of some limitations and caveats. First, attitudes, like most thinking process, are complex and context dependent. But they may fluctuate more widely than narrower content knowledge topics. The attitudes toward learning that students bring to our classroom may vary from day-to-day, depending on everything from whether they attended a party instead of studying the night before to whether a professor in another class has given a difficult and time-consuming homework assignment. Second, students' understanding of their own functional attitudes may be limited. Surveys of attitudes only measure what students think they think. To see how they really think, we have to observe them in action.

The MPEX

We created the Maryland Physics Expectations (MPEX) survey in the mid-1990s to provide a survey that could give some measure of what was happening to our students along the dimensions of relevance to the hidden curriculum [Redish 1998]. The focus of the survey was not on students' attitudes in general, such as their epistemologies or beliefs about the nature of science and scientific knowledge, but rather on their *expectations*. By expectations we mean that we want the students to ask themselves: "What do I expect to have to do in order to

[12] See chapter 3 for more discussion of the hidden curriculum.

TABLE 5.1 The Items of the MPEX Reality Cluster.

#10:	*Physical laws have little relation to what I experience in the real world.* (D)
#18:	*To understand physics, I sometimes think about my personal experiences and relate them to the topic being analyzed.* (A)
#22:	*Physics is related to the real world and it sometimes helps to think about the connection, but it is rarely essential for what I have to do in this course.* (D)
#25:	*Learning physics helps me understand situations in my everyday life.* (A)

succeed in this class?" I emphasize the narrowness of this goal: "this class," not "all my science classes" or "school in general."

The MPEX consists of 34 statements with which the students are asked to agree or disagree on a 5-point scale, from strongly agree to strongly disagree.[13] The MPEX items were validated through approximately 100 hours of interviews, listening to students talk about each item, how they interpreted it, and why they chose the answer they did. In addition, the parity of the favorable MPEX responses was validated by offering it to a series of expert physics instructors and asking what answers they would want their students to give on each item [Redish 1998]. The desired parity (agree or disagree) is labeled the *favorable* response, and the undesired parity is labeled *unfavorable.*

To illustrate the MPEX focus on expectations, consider the items given in Table 5.1. The favorable response (agree = A or disagree = D) is indicated at the end of the item. These items ask students to evaluate the link between physics and their everyday experience in two ways: from the class to their outside experience and from their outside experience to the class. Each direction is represented by two elements: one to which the favorable response is positive and one to which it is negative. The more general item "Physics has to do with what happens in the real world" was omitted, since almost all students agreed with it.

MPEX results

We analyze and display the results on the MPEX by creating an *agree/disagree (A/D) plot.* (See Figure 5.7.) In this plot, "agree" and "strongly agree" are merged (into "A"), and "disagree" and "strongly disagree" are merged (into "D"). The result is a three-point scale: agree, neutral, and disagree. This collapse of scale is based on the idea that, while it might be difficult to compare one student's "strongly agree" to another student's "agree," or to make much of a shift from "strongly agree" to "agree" in a single student, there is a robust difference between "agree" and "disagree" and a shift from one to the other is significant. The unfavorable responses are plotted on the abscissa and the favorable responses on the ordinate. Since the A + D + N ("neutral") responses must add to 100%, the point representing a class lies within the triangle bounded by the abscissa (unfavorable axis), the ordinate (favorable axis), and the line representing 0% neutral choices (F + U = 100). The expert responses are plotted as a cross in the favorable corner of the triangle.

[13] In education circles, such a ranking is referred to as a *Likert* (Lĭk-ert) scale.

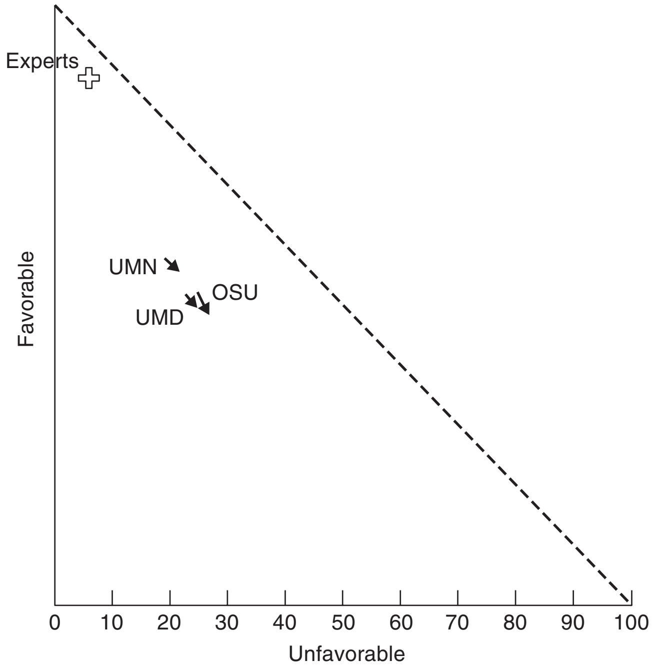


Figure 5.7 An agree/disagree (A/D) plot of overall MPEX results in large university settings at the beginning and end of a one-semester calculus-based physics class. All classes had traditional lectures with a one-hour active-engagement small-class session per week. Each represents data from ~500 students [Redish 1998].

Pre- and post-scores can be plotted for each cluster on the A-D plot. It is convenient to connect the pre- and post-results for each cluster with an arrow. A spreadsheet that allows you to paste in your MPEX results and generate A-D plots is included on the Resource CD associated with the volume.[14] Overall data from three large state universities are shown in Figure 5.7 (data from [Redish 1998]).

The MPEX has been delivered as a pre-post survey to thousands of students around the United States. The results have been remarkably consistent.

1. On the average, college and university students entering calculus-based physics classes choose favorable response on approximately 65% of the MPEX items.
2. At the end of one semester of instruction in large lecture classes, the number of favorable responses drops by approximately 1.5σ. This is true even in classes that contain active-engagement elements that produce significantly improved conceptual gains as measured, say, by the FCI.

[14] This spreadsheet was created by Jeff Saul and Michael Wittmann.

Analyzing the MPEX

Among the 34 items of the MPEX, 21 are associated into five clusters corresponding to the three Hammer variables described in chapter 3, plus two more. The five MPEX clusters are described in Table 5.2.

Note that some of the MPEX items belong to more than one cluster. This is because the MPEX variables are not interpreted as being linearly independent. The breakdown into clusters of the MPEX results at Maryland in the first semester of engineering physics is shown in Figure 5.8. These results are averaged over seven instructors and represent a total of 445 students (matched data). We see that there are significant losses in all of the clusters except concepts where we had a small gain. Note that not all of the MPEX items have been assigned to clusters.

Items 3, 6, 7, 24, and 31 (an "Effort cluster") are included to help an instructor understand what the students actually do in the class. They include items such as

> #7: *I read the text in detail and work through many of the examples given there.* (A)
>
> #31: *I use the mistakes I make on homework and on exam problems as clues to what I need to do to understand the material better.* (A)

Although students' answers to these items are interesting, I recommend that they not be included in overall pre-post analyses. There is a strong tendency for students to hope that they will do these things before a class begins, but they report that they didn't actually do them after the class is over. Inclusion of these items biases the overall results toward the negative.

MPEX items 1, 5, 9, 11, 28, 30, 32, 33, and 34 are not assigned to clusters. Interviews suggest that these items are indeed correlated with student sophistication, but they do not correlate nicely into clusters. Furthermore, since the MPEX was designed for a class in calculus-based physics for engineers, some of these items may not be considered as desirable goals for other classes.

Table 5.2 The MPEX Variables and the Assignment of Elements to Clusters.

	Favorable	Unfavorable	MPEX Items
Independence	Takes responsibility for constructing own understanding	Takes what is given by authorities (teacher, text) without evaluation	8, 13, 14, 17, 27
Coherence	Believes physics needs to be considered as a connected, consistent framework	Believes physics can be treated as unrelated facts or independent "pieces"	12, 15, 16, 21, 29
Concepts	Stresses understanding of the underlying ideas and concepts	Focuses on memorizing and using formulas without interpretation or "sense-making"	4, 14, 19, 23, 26, 27
Reality	Believes ideas learned in physics are relevant and useful in a wide variety of real contexts	Believes ideas learned in physics are unrelated to experiences outside the classroom	10, 18, 22, 25
Math link	Considers mathematics as a convenient way of representing physical phenomena	Views the physics and math independently with little relationship between them	2, 8, 15, 16, 17, 20

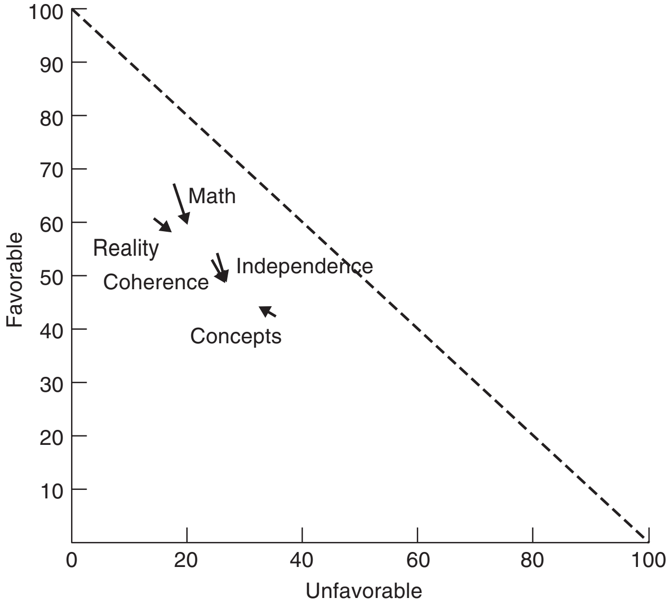


Figure 5.8 Pre-post shifts on the MPEX clusters at the University of Maryland in the first semester of engineering physics (data from [Redish 1998]).

Two items in particular tend to be controversial.

> #1: *All I need to do to understand most of the basic ideas in this course is just read the text, work most of the problems, and/or pay close attention in class.* (D)
>
> #34: *Learning physics requires that I substantially rethink, restructure, and reorganize the information that I am given in class and/or in the text.* (A)

Sagredo is unhappy about these. He says, "For #1, I would be happy if they did that. Why do you want them to disagree? For #34, some of my best students don't have to do this to do very well in my class. Why should they agree?" You are correct, Sagredo, and I suggest that you give these items, but not include them in your analysis or plots.[15] We include them because our interviews have revealed that the best and most sophisticated students in a class who are working deeply with the physics respond favorably as indicated. Certainly, for introductory courses this level of sophistication may be unnecessary. I like to retain these items, hoping that something I am doing is helping my students realize that developing a deep understanding of physics requires the responses as indicated.

Getting improvements on the MPEX

The fact that most courses probed with the MPEX show losses is discouraging but not unexpected. It is not surprising that students do not learn elements of the hidden curriculum

[15] This is easily achieved in the Saul-Wittmann MPEX analysis spreadsheet by replacing "1"s by "0"s in the appropriate cells.

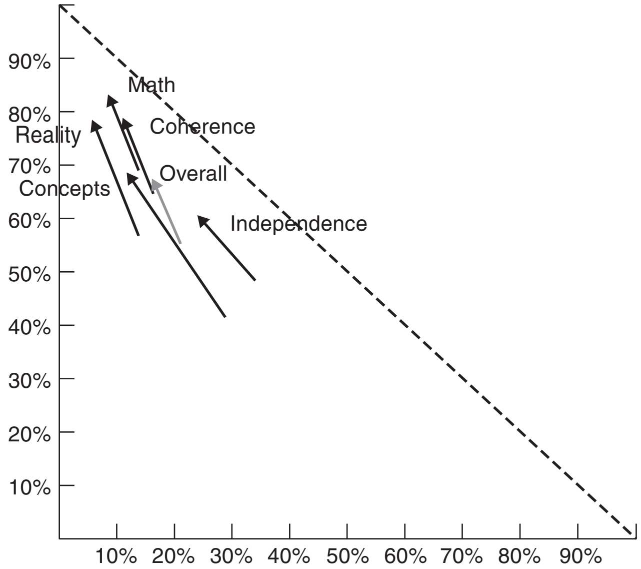


Figure 5.9 An A/D plot for pre-post MPEX results for Andy Elby's physics class at Thomas Jefferson High School in Virginia. For each cluster, the pre-result is at the base of the arrow, the post is at the tip of the arrow, and the name of the associated cluster is next to the arrowhead. The overall result is shown as a gray arrow [Elby 2001].

as long as it stays hidden. If we want students to improve along these dimensions, we have to be more explicit in providing structures to help them learn them.

Recently, MPEX results in classes designed to focus on explicit instruction in intuition building, coherence, and self-awareness of one's physics thinking have shown substantial improvements in all the MPEX categories [Elby 2001]. These results are shown in Figure 5.9.

I have been able to achieve MPEX gains in my large lecture classes by making specific efforts to keep issues of process explicit in both lectures and homework. Soon after developing the MPEX in 1995, I made strong efforts in my calculus-based physics classes to produce gains by giving estimation problems to encourage a reality link, by talking about process, and by stressing interpretation of equations in lecture. Results were disappointing. The responses on the reality link items still deteriorated, as did overall results. After much thought and effort, I introduced activities in lecture to help students become more engaged in these issues (see the discussion of Interactive Lecture Demonstrations in chapter 8), and I expanded my homework problems to include context-related problems every week. I reduced the number of equations I used and stressed derivations and the complex application of the few conceptually oriented equations that remained. The results (in my algebra-based class in 2000) were the first MPEX gains I had ever been able to realize.[16] Some of the results for four interesting items are shown in Table 5.3.

[16] This class also produced the largest FCI/FMCE gains I have ever managed to achieve.

TABLE 5.3 Pre- and Post-results on Four MPEX Items from a Calculus-Based Class Using UW Tutorials and Algebra-Based Class Using More Explicit Self-Analysis Techniques.

			Calculus-based (1995)			Algebra-based (2000)		
			F	U	N	F	U	N
#4	"Problem solving" in physics basically means matching problems with facts or equations and then substituting values to get a number.	Pre	60%	21%	19%	66%	30%	4%
		Post	77%	13%	10%	91%	9%	0%
#13	My grade in this course is primarily determined by how familiar I am with the material. Insight or creativity has little to do with it.	Pre	54%	24%	22%	57%	40%	3%
		Post	49%	23%	28%	79%	19%	2%
#14	Learning physics is a matter of acquiring knowledge that is specifically located in the laws, principles, and equations given in class and/or in the textbook.	Pre	39%	28%	33%	36%	53%	11%
		Post	37%	24%	39%	64%	34%	2%
(#19)	The most crucial thing in solving a physics problem is finding the right equation to use.	Pre	43%	32%	25%	45%	45%	10%
		Post	46%	26%	28%	72%	26%	2%

The MPEX serves as a sort of "canary in the mine" to detect classes that may be toxic to our hidden curriculum goals. The fact that most first-semester physics classes result in a deterioration of favorable results is telling. The fact that gains can be obtained by strong and carefully thought out efforts suggests that the use of the MPEX can be instructive, when judiciously applied.

The VASS

A second survey on student attitudes toward science was developed by Ibrahim Halloun and David Hestenes [Halloun 1996]. The Views about Science Survey (VASS) comes in four forms: one each for physics, chemistry, biology, and mathematics. The physics survey has 30 items. Each item offers two responses, and students respond to each item on an eight-point scale as shown in Figure 5.10. (Option 8 is rarely chosen.) In addition to items that probe what I have called expectations, the survey includes items that attempt to probe a student's epistemological stance toward science. A sample item is given in Figure 5.11.

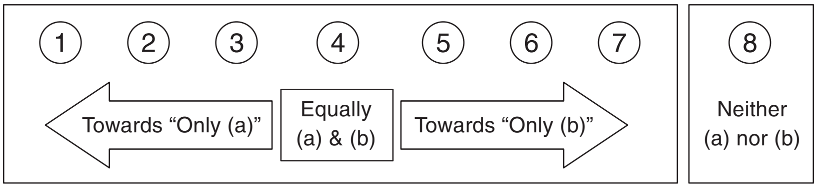


Figure 5.10 The eight-point scale for responding to items from the VASS [Halloun 1996].

The VASS is designed to probe student characteristics on six attitudinal dimensions—three scientific and three cognitive.

Scientific dimensions of the VASS

1. *Structure of scientific knowledge:* Science is a coherent body of knowledge about patterns in nature revealed by careful investigation rather than a loose collection of directly perceived facts (comparable to MPEX coherence cluster).
2. *Methodology of science:* The methods of science are systematic and generic rather than idiosyncratic and situation specific; mathematical modeling for problem solving involves more than selecting mathematical formulas for number crunching (extends MPEX math cluster).
3. *Approximate validity of scientific results:* Scientific knowledge is approximate, tentative, and refutable rather than exact, absolute, and final (not covered in the MPEX).

Cognitive dimensions of the VASS

4. *Learnability:* Science is learnable by anyone willing to make the effort, not just by a few talented people, and achievement depends more on personal effort than on the influence of teacher or textbook.
5. *Reflective thinking:* For a meaningful understanding of science one needs to concentrate on principles rather than just collect facts, look at things in a variety of ways, and analyze and refine one's own thinking.
6. *Personal relevance:* Science is relevant to everyone's life; it is not of exclusive concern to scientists (relates in part to MPEX reality cluster).

The favorable polarization of the VASS responses was determined by having it filled out by physics teachers and professors. Teachers' responses were strongly polarized on most items.

The laws of physics are:
(a) inherent in the nature of things and independent of how humans think.
(b) invented by physicists to organize their knowledge about the natural world.

Figure 5.11 A sample item from the VASS [Halloun 1996].

TABLE 5.4 Categories for Classifying VASS Responses [Halloun 1996].

Profile Type	Number of Items out of 30
Expert	19 or more items with expert views
High Transitional	15–18 items with expert views
Low Transitional	11–14 items with expert views and an equal or small number with folk views
Folk	11–14 items with expert views but a larger number of items with folk views or 10 items or less with expert views

Answers agreeing with the teachers are called *expert views,* while the polar opposites are referred to as *folk vie*ws. Halloun and Hestenes classify students into four categories depending how well their responses agree with those of experts (see Table 5.4).

Halloun and Hestenes delivered the VASS to over 1500 high school physics students in 39 schools (30 of which used traditional rather than active engagement methods) at the beginning of class. They found them to be classified about 10% expert, about 25% high transitional, about 35% low transitional, and about 30% folk. Surveys of beginning college physics students gave similar results. For the high school students, there was a significant correlation between the students' profiles on the VASS and their gains on the FCI, as shown in Figure 5.12.

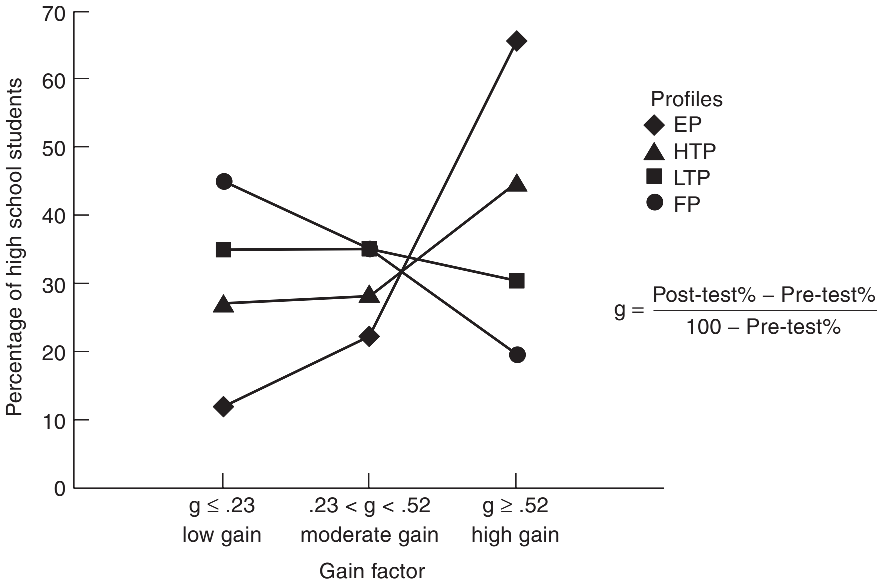


Figure 5.12 Correlation between VASS profiles and student gains on the FCI [Halloun 1996].

The EBAPS

Both the MPEX and the VASS suffer from the problem of probing what students think they think rather than how they function. In addition, they have the problem that for many items, the "answer the teacher wants" is reasonably clear, and students might choose those answers even if that's not what they believe. In the Epistemological Beliefs Assessment for Physics Science (EBAPS), Elby, Fredericksen, and White attempt to overcome these problems by presenting a mix of formats, including Likert-scale items, multiple-choice items, and "debate" items. Many EBAPS items attempt to provide context-based questions that ask students what they would do rather than what they think. The debate items are particularly interesting. Here's one.

> #26:
>
> **Justin:** When I'm learning science concepts for a test, I like to put things in my own words, so that they make sense to me.
>
> **Dave:** But putting things in your own words doesn't help you learn. The textbook was written by people who know science really well. You should learn things the way the textbook presents them.
>
> (a) I agree almost entirely with Justin.
>
> (b) Although I agree more with Justin, I think Dave makes some good points.
>
> (c) I agree (or disagree) equally with Justin and Dave.
>
> (d) Although I agree more with Dave, I think Justin makes some good points.
>
> (e) I agree almost entirely with Dave.

The EBAPS contains 17 agree-disagree items on a five-point scale, six multiple-choice items, and seven debate items for a total of 30. The Resource CD includes the EBAPS, a description of the motivations behind the EBAPS, and an Excel template for analyzing the results along five axes:

Axis 1 = Structure of knowledge
Axis 2 = Nature of learning
Axis 3 = Real-life applicability
Axis 4 = Evolving knowledge
Axis 5 = Source of ability to learn