Chapter 4. Multiple Choice Concept Tests: The Force Concept Inventory (FCI)

I. CHAPTER OVERVIEW

In the early 1980s, McDermott,[1] Viennot,[2] and other physics education researchers[3, 4] found that each student comes into a physics course not as a blank slate but brings into the classroom a system of common sense beliefs and intuitions about how the world works derived from extensive personal experience (this is discussed in more detail in chapter 2). Furthermore, they found that these common sense beliefs were very stable and often incompatible with the physics being taught in the introductory course. Traditional instruction does little to change these beliefs and they cause some students to misinterpret the course material. This result was pieced together by researchers who studied student understanding of isolated concepts in mechanics.

At the same time Hestenes and Halloun[5] at Arizona State University began developing an instrument called the Mechanics Diagnostic Test (MDT) that measured not the student's initial knowledge of Newtonian force but the discrepancy between the students' common sense beliefs and their belief in the Newtonian force concept. In 1992, an improved version of the MDT was published as the Force Concept Inventory (FCI).[6] Because the questions are written in plain language and easily understood by non-physics students, the FCI can be given at the beginning and as well as at the end of a course to see if the students improve. Because the instrument was included in the published study, which also described how to interpret the results, the FCI could be used by any physics instructor to evaluate their own students. Many physics faculty, including Eric Mazur as described in chapter 2, have overcome the initial "not my class" response to reports of students' difficulties with conceptual understanding after using the MDT and/or the FCI and seeing exactly how poorly their own students fare. The value of these two instruments has led to the development of other multiple-choice concept tests in mechanics and other content areas of the introductory course.[7] One of the other mechanics tests is the Force Motion Concept Evaluation (FMCE),[8] an instrument similar to the FCI that looks at a smaller set of concepts and makes heavy use of graphical and pictorial representations.

The FCI and the FMCE are the two most commonly used physics concept tests in use today. Every class involved in this study has used one of these two tests to monitor students' improvements in conceptual understanding. Almost all evaluations of improved introductory mechanics classes reported at national AAPT meetings make use of one of these tests. Yet, recently questions have been raised concerning what tests like the FCI actually measure.[9,10]

This chapter will review the development of these tests, discuss what they measure, and present some results concerning limitations of the FCI and how the two tests compare. The so-called 4-H controversy will also be discussed.[11, 12, 13, 14]

II. DEVELOPMENT OF THE FCI

Halloun and Hestenes developed the MDT for the expressed purpose of evaluating introductory physics instruction objectively. They share Reif's view[15] that student learning should be viewed as a transformation from an initial state to a final state. Thus, the MDT was intended to be used as a pre-course test to assess students' initial knowledge state and as a post-course test to measure the effect of instruction independent of other assessments such as exams or homework. Because Newtonian mechanics is the central theme of the first course in most introductory physics sequences and it is an essential for the rest of the sequence, Halloun and Hestenes restricted their instrument to concepts related to Newtonian force.

Halloun and Hestenes developed the MDT in successive iterations over three years and administered it to over 1000 students in introductory college physics. The test questions were initially selected to assess students' qualitative conceptions of motion and its causes. The questions were designed to identify common misconceptions or non-Newtonian common sense beliefs noted in the literature. The early versions required written responses. Later multiple-choice versions used the most common written student responses that were indicative of non-Newtonian common sense beliefs as distractors. A student's overall score was taken as a measure of that student's qualitative understanding of Newtonian force.

The FCI is similar in design to the MDT and produces similar overall scores when used in comparable classes. In fact, roughly half the questions from the MDT remain unchanged in the FCI. The main advance in the FCI from the MDT comes from a systematic analysis and explicit taxonomy of the basic concepts of Newtonian concepts and students' common sense beliefs. The FCI is designed to cover these concepts more comprehensively and facilitate the interpretation of the results. For example, the FCI can identify students' difficulties with each of Newton's laws of motion and can help identify the common sense beliefs associated with each of these difficulties.[16]

Table 4-1. Newtonian Concepts in the Force Concept Inventory. First appeared in D. Hestenes, M. Wells, and G. Swackhamer, "Force Concept Inventory", in the March 1992 Physics Teacher.[17]

Concept	Inventory Item
0. Kinematics	
Velocity discriminated from position	20E
Acceleration discriminated from velocity	21D
Constant acceleration entails	
parabolic orbit	23D;24E
changing speed	25B
Vector addition of velocities (7E)	
1. First Law	
with no force	4B;(6B);10B
velocity direction constant	26B
speed constant	8A;27A
with canceling forces	18B;28C
2. Second Law	
Impulsive force	(6B);(7E)
Constant force implies constant acceleration	24E;25B
3. Third Law	
for impulsive forces	2E;11E
for continuous forces	13A;14A
4. Superposition Principle	
Vector sum	19B
Canceling forces	(9D);18B;28C
5. Kinds of Force	
5S. Solid contact	
passive	(9D);(12 B,D)
Impulsive	15C
Friction opposes motion	29C
5F. Fluid contact	
Air resistance	22D
buoyant (air pressure)	12D
5G. Gravitation	5D;9D;(12 B,D);17C;18B;22D
acceleration independent of weight	1C;3A
parabolic trajectory	16B;23D

Table 4-2. Taxonomy of misconceptions probed by the Force Concept Inventory (FCI). Belief in the misconceptions is suggested by selection of the corresponding FCI item.[17]

Misconception	Inventory Item
0. Kinematics	
K1. Position-velocity undiscriminated	20B,C,D
K2. velocity-acceleration undiscriminated	20A;21B,C
K3. nonvectorial velocity composition	7C
1. Impetus	
I1. impetus supplied by 'hit'	9B,C;22B,C,E;29D
I2. loss/recovery of original impetus	4D;6CE;24A;26A,D,E
I3. impetus dissipation	5A,B,C;8C;16C,D;23E; 27C,E;29B
I4. gradual/delayed impetus build-up	6D;8B,D;24D;29E
I5. circular impetus	4A,D;10A
2. Active Force	
AF1. only active agents exert forces	11B;12B;13D;14D; 15A,B;18D;22A
AF2. motion implies active force	29A
AF3. no motion implies no force	12E
AF4. velocity proportional to applied force	25A;28A
AF5. acceleration implies increasing force	17B
AF6. force causes acceleration to terminal velocity	17A;25D
AF7. active force wears out	25C,E
3. Action/Reaction Pairs	
AR1. greater mass implies greater force	2A,D;11D;13B;14B
AR2. most active agent produces greatest force	13C;11D;14C
4. Concatenation of Influences	
CI1 largest force determines motion	18A,E;19A
CI2. force compromise determines motion	4C;10D;16A;19C,D;23C;24C
CI3 last force to act determines motion	6A;7B;24B;26C
5. Other Influences on Motion	
CF. Centrifugal force	4C,D,E;10C,D,E
Ob. Obstacles exert no force	2C;9A,B;12A;13E;14E
Resistance	
R1. mass makes things stop	29A,B;23A,B
R2. motion when force overcomes resistance	28B,D
R3. resistance opposes force/impetus	28E
Gravity	
G1. air pressure-assisted gravity	9A;12C;17E;18E
G2. gravity intrinsic to mass	5E;9E;17D
G3. heavier objects fall faster	1A;3B,D
G4. gravity increases as objects fall	5B;17B
G5. gravity acts after impetus wears down	5B;16D;23E

III. EVALUATION OF FCI RESULTS

Unlike most multiple choice tests, the FCI and the earlier MDT distractors come from common-sense student responses to open-ended versions of the questions. This forces the student to choose between the Newtonian concepts and the common sense alternatives. To help interpret FCI results, Hestenes *et al*. have included two taxonomy tables, one for the Newtonian force concept (Table 4-1) and one for the common sense alternatives (Table 4-2).[18]

In Table 4-1, the Newtonian force concept is decomposed into six conceptual dimensions: kinematics, first law, second law, third law, superposition principle, and kinds of force. Most physics instructors will agree that all six dimensions are required for a complete understanding of the Newtonian force concept. Each dimension is further broken down into the isolated concepts that characterize that dimension. Each concept is listed with the FCI items in which it appears and the associated Newtonian response. Note that with the exception of item 12, there is only one Newtonian response to each item.[19] Although each dimension is probed by questions of more than one type, most of the concepts in each dimension are only probed by one or two questions. This is an important point we will come back to at the end of this chapter.

IV. VALIDATION AND RELIABILITY OF THE MDT & THE FCI

A. The MDT

Content validity establishes that the items or questions are relevant to what is being studied and that the subjects' interpretation of both the questions and answers matches what the developers intended. The content validity of the MDT was established in four different ways. First, early versions of the test were critiqued by physics professors and physics graduate students. Second, the MDT was taken by graduate students to verify agreement on the correct answers. Third, interviews conducted with introductory physics students verified that they understood the questions and the multiple-choice responses. Fourth, the responses of thirty-one students who received A's in the introductory physics class were checked for common misunderstandings of individual items.

Reliability indicates that the instrument results or the results from a subsection of the instrument are reproducible for a given group of subjects. The reliability of the MDT was established through interviews and a statistical analysis. A sample of students who had already taken the test were interviewed. During the interviews, the students repeated their responses on the MDT "virtually without exception." The students gave reasons for their choices and were not easily swayed from their answers as the individual items were discussed. This is an indication that the students' interview responses reflect stable beliefs as opposed to tentative, random, or flippant responses.

The statistical analysis used the Kuder-Richardson formula, otherwise known as the KR-20.[20] The KR-20 is a special case of the Cronbach Alpha coefficient (see the more detailed discussion on reliability in chapter 5). Both are methods for determining the internal consistency of the test items. Both methods are measures of the intercorrelation of the test items, that is, how well the responses to the items in the test contribute to the overall test score. The KR-20 is used for tests where the items are scored right or wrong. The KR-20 was used on pre-course and post-course MDT results from different, but comparable groups. The KR-20 values obtained were 0.86 for the pre-course results and 0.89 for the post-course results. A test with a KR-20 value greater than 0.70 is considered reliable enough for measurements of groups. A test with a KR-20 greater than 0.80 is considered reliable enough for measurement of individuals.[21] The KR-20 values for the MDT indicate high reliability and suggest that even measurements of individuals should be considered reliable.

B. The FCI

Hestenes, Wells, and Swackhamer did not follow the formal procedures to establish the validity or reliability of the FCI because the FCI is not substantially different from the MDT whose validity and reliability had been established as discussed above. They back up this claim by noting that the FCI and the MDT share a common design, that 14 of the 29 questions on the FCI came unchanged from the MDT, and the results of the FCI and MDT on comparable classes are very similar.

However, Hestenes *et al*. did interview 20 high school students and 16 first-year graduate students about their responses to the FCI. The responses of the high school students were very predictable and many of the responses sounded as if the students were reciting the results of the studies that led to publication of the MDT.[22] They had firm reasons for most of their choices, although there was some vacillation among alternatives. Few non-Newtonian responses were made by students who understand the Newtonian concept involved (false negatives), but Newtonian responses without Newtonian reasoning were “fairly common” (false positive). Because of the false positives and the relatively few false negatives, Hestenes *et al*. state that except for high score (> 80% or so) the FCI score “should be regarded as an upper bound on a student’s Newtonian understanding.”[23]

The 16 first year graduate students were beginning the graduate mechanics course at Arizona State University. They participated in in-depth interviews on the questions they missed on the inventory. Although some of the difficulties were attributed to language difficulties both with the foreign students and the native English speakers, the interviews confirmed difficulties with Newton's 2^{nd} and 3^{rd} laws, buoyancy, and friction that had been indicated by the FCI.

In addition, Hestenes *et al*. demonstrated the reproducibility of the FCI, the mark of a good experimental measurement. It is difficult to give the FCI to the same students twice over a period of time long enough for the students not to remember the questions but short enough that they haven't learned anything or changed the way they think about forces. But, if the FCI is a reliable test, then classes with similar populations (comparable classes) should get comparable FCI scores. Hestenes *et al.* found that the post-course averages from classes with over a thousand students in introductory classes taught in the traditional style by seven different professors at Arizona State University were remarkably similar. The post-course averages were all between 60 and 64% correct. The reproducibility of FCI results from comparable classes has since been observed by others including Hake (see Hake's 6000 Student Study and the h-factor discussion in section V. below) and myself (see chapter 9)

More recently, Hestenes and Halloun[24] have used interviews with students and other tests to compare students FCI scores to other measures of Newtonian skills including problem solving.[25] Based on their results, Hestenes and Halloun suggest an interpretation of FCI scores that is consistent with a three-stage model of conceptual understanding in learning Newtonian mechanics.

Students who score below 60% on the FCI are classified as stage I. Stage I student thinking can typically be described in terms of the following characteristics:

- undifferentiated concepts of velocity and acceleration,
- lack of a vectorial concept of velocity,
- belief that there are other influences on motion besides forces,
- inability to reliably identify passive and active agents of force on an object, and
- fragmented and incoherent concepts about force and motion.

Students who score between 60% and 85% on the FCI are in stage II. Hestenes and Halloun suggest that an FCI score of 60% be considered as the entry threshold to Newtonian thinking. In stage II, students are developing coherent dynamics' concepts, including vectorial concepts of velocity, acceleration, and force. An FCI score of 85% is interpreted as the threshold to stage III and mastery of the Newtonian force concept. Students in stage III develop a complete Newtonian interaction concept including a full understanding of Newton's third law. Hestenes and Halloun express confidence in "identifying students with scores above [85%] as confirmed Newtonian thinkers."[26]

In addition, Hestenes and Halloun believe their results indicate that not only is the conceptual development of students influenced by the order in which the concepts are introduced in instruction, but that these is a natural order in which the concept are most easily learned. A similar study of student understanding of kinematics by Thornton also suggests the existence of a natural order for learning mechanics concepts.[27]

V. RESULTS FROM FCI/MDT STUDIES

A. Halloun, Hestenes, and the MDT (1985)

Using the MDT in conjunction with a math-skills diagnostic test of their own design, Halloun and Hestenes found that the initial knowledge state of the students as measured by the two pretests is a significant predictor of student performance in traditional introductory physics classes.[28] In combination the two tests accounted for about 42% of the variance of exam scores,[29] that is 42% of the variability in exams scores can be predicted by the two diagnostic tests. The MDT alone accounts for 31% of the variance. Differences in students' gender, age, academic major, and academic background were found to have little or no effect. They also found that the pretest MDT scores for calculus-based classes (51%, 51%, 50%, and 53%) were significantly better than the scores for algebra/trig-based classes (37%, 37%, 40%) with relatively narrow distributions for both. That the calculus-based classes did better is not surprising, but considering the simplicity of the tests, the math and physics diagnostics average scores were thought to be low for both types of classes. The reader is reminded that a low score on the MDT does not just mean that students don't know or understand the concepts of Newtonian mechanics, but that common sense misconceptions may be firmly in place. Based on my own study of FCI scores from both calculus-based and algebra/trig-based courses from 17 colleges and universities, the scores Halloun and Hestenes found in 1985 are typical pre-course scores for calculus-based and algebra/trig-based courses for schools of moderate selectivity.

The teaching styles of the four instructors in the calculus-based introductory classes in the Halloun and Hestenes study were very different. Professor A was a theorist whose lectures emphasized the conceptual structure of physics with careful definitions and orderly logical arguments. Professor B expended a great deal of time and energy to incorporate many demonstrations in his lectures to help students develop physical intuition. Professor C emphasized problem solving and modeled one problem after another in his lectures. Professor D as an experimental physicist teaching physics for the first time tended to follow the book closely in his lectures. All four professors were known as good teachers according to informal peer opinion and student evaluation. In fact, Professor B won two awards for outstanding teaching. The students of professors A, B, and C increased their scores on the MDT by 13%; the students of professor D increased by 11%.

Despite the wide variety of teaching styles of the four instructors, the result on the MDT was essentially the same. Halloun and Hestenes therefore conclude from this result “that the basic knowledge gain under traditional lecture instruction is essentially independent of the professor.” They comment that this result is consistent with observations by physics instructors that “the most strenuous efforts to improve instruction hardly seem to have any effect on the general student performance.”

Halloun and Hestenes interpret these results in the following way. The low initial level and small gain of students basic knowledge as measured by the MDT indicate that throughout the course students are operating with a defective vocabulary and belief system. This implies that students are continually misunderstanding the course material presented in class. Furthermore, in their words,

> The small gains in basic knowledge ‘explain’ the high predictive validity of the mechanics pretest; the student’s ability to process information in the course depends mainly on his initial knowledge state and hardly

> improves throughout the course. The high predictive [power] of the [MDT] pretest is not intrinsic to the test; rather, it indicates a failure of [traditional] instruction. The more effective the instruction is in altering the basic knowledge state [of the students], the lower the predictive [power] of the pretest.[30]

B. The Hestenes, Wells, and Swackhamer Study (1992)

Hestenes, Wells, and Swackhamer used the FCI as a pre/post course evaluation of high school and university introductory physics classes.[31] They found low pre-course FCI scores for most classes and smaller post-course FCI scores for traditional classes. The classes using research-based pedagogy tended to have higher post-test scores.

Their study of high school classes included 19 high school teachers involved in an NSF physics instruction enhancement program run by Wells and Hestenes. The FCI was used as a pre/post evaluation of the high school teachers' classes. Hestenes *et al.* compared the FCI scores with ratings of the socioeconomic level of the students at the teacher's schools and with a ranking of the teacher's competence as measured by academic background, concept test score, and teaching experience. The results appeared independent of socioeconomic level of the school but this could be due to selection bias since the students who take physics tend not to be typical of the student population at the school. With one exception, the results tend to be independent of the teacher competence ranking. One instructor scored 39% on the Mechanics Diagnostic; that instructor's students only scored 33% on the post-course test, the lowest post score in the study. This implies that the students' FCI scores may only be affected by the quality of the instructor in extreme cases.

As part of the enhancement program, the teachers took a workshop on the teaching methods of Wells. The method uses computer-assisted laboratory-oriented instruction with no lectures, but with substantial class discussion. The high school teachers' classes were tested before and after the workshop. The classes showed little or no improvement. Upon further examination, Hestenes *et al*. found that in their first year of instruction with the Wells' method the teachers focused on the mechanics of the method and not on the pedagogy. This suggests that technology by itself cannot improve instruction. The best that technology can do is enhance the effectiveness of good pedagogy.

C. Hake's 6000 Student Study and the h-factor (1993-97)

A detailed study of pre/post MDT and FCI results nationwide by Hake compares the performance of 62 introductory physics classes (N = 6542 students).[32] The student populations sampled include high schools, colleges, and universities. Hake's results 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). This is shown schematically in the Hake plot shown in Fig. 4-1.[33] Traditional classes lie on the line closest to the horizontal axis and show limited improvement. The middle line represents classes with active engagement (see chapter 2 for a discussion of active engagement). The steepest line represents classes with active engagement and a research-based text. The negative slope of the line from a data point to the point (100,0) is a figure of merit:

h = (class post-test average - class pre-test average)/(100 - class pre-test average)

The interpretation of this is that two classes having the same figure of merit, *h*, have achieved the same fraction of the possible gain.

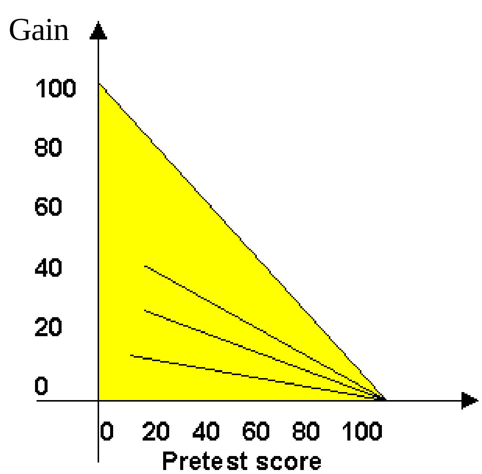
The 62 classes are classified as either "traditional" or "active engagement."

These terms are defined as follows:

Active Engagement: Courses that incorporate at least in part research-based teaching methods where the students are actively involved in discussion, discovery and/or analysis of the course material. The students work on these activities in cooperative groups or for at least one hour per week in class and in some cases as many as six hours per week. In these active learning activities, the teaching emphasis is on helping the students answer their own questions and think about what they are learning.

Traditional: Courses that do not make use of active engagement activities. Typically, the primary activity in the lecture and recitation components of a traditional class involves students listening to a lecture presentation or answers

Figure 4-1. Schematic of the Hake plot. A class's average pre-test FCI score is plotted on the horizontal axis, the pre- to post-test gain is plotted on the vertical axis. Since the maximum class average is 100%, every data point must lie in the shaded region. The three lines of constant fractional gain (*h*) described in the text are shown in the figure below. (This figure is reproduced from "On the effectiveness of active-engagement microcomputer-based laboratories" by Redish, Saul, and Steinberg in the American Journal of Physics.[34])



to their questions by an instructor. Students may make predictions for demonstrations but do not discuss them with their peers or write them down. Hands-on activities like laboratories follow fairly detailed instructions that minimize decision making.

The main distinction of an active engagement activity is whether or not a majority of the students are actively discussing, arguing, or testing the course material. The classes in Hake's study were categorized based on information from the instructors either through private communications with Hake or through published accounts of the curriculum. Hake found that active engagement classes had significantly higher fractional gain than traditional classes. A Hake plot of the FCI/MDT scores from his survey is shown in Figure 4-2. The average fractional gains for both types of classes are:

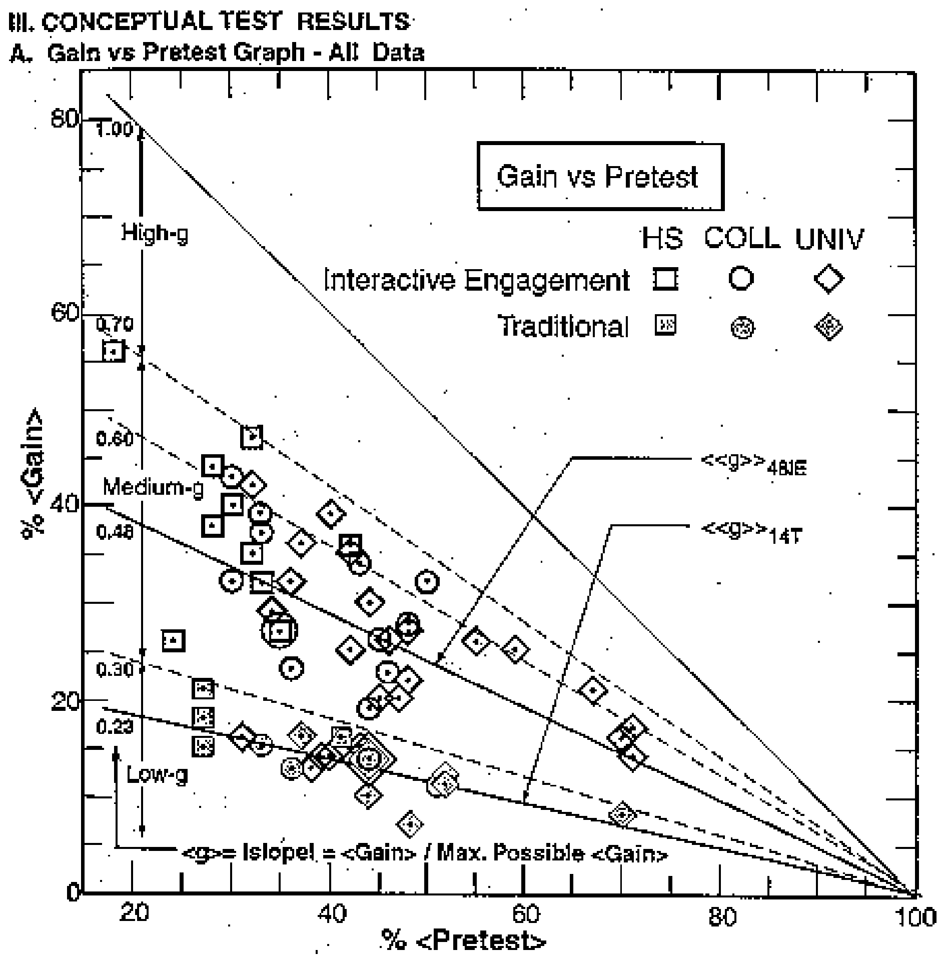
Traditional Classes (14 classes) <*h*> = 0.23 ± 0.04 (std. dev.)

Active Engagement Classes (48 classes) <*h*> = 0.48 ± 0.14 (std. dev.)

where *h* is averaged over classes, not students.

To test the significance of this result, I performed a two tailed t-test with pooled variance on Hake's data.[35] Assuming that these two measurements are drawn from similar populations, the difference between the two types of classes is statistically significant since the probability that this difference in means is due to random chance is less than 0.1 %. This result clearly shows that the active engagement classes are achieving significantly greater fractions of the possible gain on the FCI. All but seven of the forty-eight active engagement classes had higher *h*'s than all of the traditional classes. Hake noted that case studies showed that these 7 active engagement classes with relatively low <*h*>'s (0.20 < <*h*> < 0.28) had various implementation problems

Figure 4-2. Hake Plot from 6000 student study of MD and FCI results.[36] Percent gain is plotted vs. percent average pretest score. The data shown below represents pre- and post-test results from 62 courses, 14 traditional lecture courses (N = 2084 students) and 48 active engagement classes (N = 4458 students). The slope line labeled $\langle\langle g\rangle\rangle_{48IE}$ represents the average fractional gain *h* for the 48 active engagement classes. The slope line labeled $\langle\langle g\rangle\rangle_{14T}$ represents the average fractional gain *h* for the 14 traditional lecture classes.



with the active engagement methods. These problems included:

> insufficient training of instructors new to [active engagement] methods, failure to communicate to students the nature of the science and learning, lack of grade incentives for taking [active engagement] activities seriously, a paucity of exam questions which probe the degree of conceptual understanding induced by [active engagement] methods, and use of active engagement in only isolated components of the course.[37]

Note that several of these problems regarding communication and assessment are related to the issues of the "hidden curriculum" discussed throughout this dissertation.

Hake's inference that *h* is a measure of course effectiveness is bolstered by examining the correlations between the average normalized gain $\langle h \rangle$, the average unnormalized gain g ($g \equiv$ average post-test score - average pre-test), and the average pre-test scores averaged over all 62 classes. There is a low correlation between *h* and the pre-course score (r = 0.02). This implies that students' knowledge coming into the course does not strongly influence normalized gain, *h*. There is also a negative correlation between the unnormalized gain (post-course score - pre-course score) and the pre-course score (r = -0.49). This negative correlation indicates that classes that start with smaller pre-course scores tend to have larger unnormalized gains than classes of the same category with larger pre-course scores. The latter correlation is quite evident from the plot of Hake's FCI/MDT survey data shown in Figure 4-2.

Hake collected this data to address the question, "Can the classroom use of [active engagement] methods increase the effectiveness of introductory mechanics courses well beyond that attained by traditional methods?" Note that this question is similar to the research questions explored by this dissertation, but Hake limits his study to the results of multiple-choice tests. (There are also design differences between the two studies discussed below.) To answer this question, Hake solicited pre/post FCI and MDT data from the physics education community through talks at colloquia and meetings and e-mail postings on the PHYS-L and PhysLrnR listserves.[38] Hake acknowledges that his method of data solicitation "tends to pre-select results which are biased in favor of outstanding courses which show relatively high gains on the FCI;" because relatively low gains are

> neither published not communicated except by those who wish to use the results from a traditional course as a baseline … or possess unusual scientific objectivity and detachment. Fortunately, several in the latter category contributed data to the present survey for courses in which active engagement methods were used but relatively low gains were achieved.[39]

However, Hake also notes that this survey includes all the pre/post FCI results from traditional classes that were known to him.

In addition, while Hake controlled or took into account several kinds of systematic error, not all the data Hake presents is matched so that the result represents real changes in the students and not differences in which students took the FCI.[40] Hake estimates that the error in the normalized gain h is probably less than 5% for classes with 20-50 students and less than 2% for classes with more than 50 students. In my own studies of FCI results reported in chapter 9, I found that the matched and unmatched results are not significantly different. Despite the known flaws in Hake's study, his result of a very significant difference in <*h*> for traditional and active engagement classes is still quite impressive.

Prior to this dissertation, Hake's study was the only one to compare concept test results from classes using a variety of active-engagement methods with traditional instruction. In chapter 9, I present FCI results from approximately 2000 students from calculus-based introductory physics classes using three research-based curricula using active-learning methods and traditional instruction from eight of the ten schools participating in this study. Where possible, results are presented from traditional classes and active engagement classes at the same school. The selection bias in Hake's data was avoided since data was taken from all first term classes in the sequence being studied at a particular school. The data are matched unless explicitly stated otherwise.

VI. DISCUSSION OF WHAT IS MEASURED BY THE FCI

A. The 4-H Controversy (1995)

The so-called 4-H controversy began in March 1995 when Huffman and Heller reported on the results of a factor analysis of FCI data.[41] They questioned what is measured by the FCI and the dimensions of the Newtonian force concepts defined by Hestenes *et al*. and shown in Table 4-1. Hestenes and Halloun defended their interpretation of the test and were in turn rebutted by Heller and Huffman.

The point of debate in this controversy is quite subtle. Hestenes, Halloun, Heller and Huffman all acknowledge that the FCI is a useful test for evaluating instruction and as a diagnostic tool for classes. All four also agree that Hestenes *et al*. have demonstrated the reliability (the results for similar classes are reproducible), the face validity (physics instructors agree that Halloun et al.'s six dimensions are necessary for a Newtonian force view and that the Newtonian choice for each item is correct), and context validity (the developers interviewed students to confirm the validity of the student responses to individual items) of the FCI. What is in dispute is the construct validity of the FCI, that is whether or not the FCI actually measures the student force concept and the six conceptual dimensions of force.

To understand the nature of the controversy, it is necessary to understand that a factor analysis is a statistical method for finding structure in a set of data by analyzing the inter-item correlations. The factors are those groups of items that show relatively strong correlations with each other. A more detailed description can be found in the next chapter in the section on reliability. Factor analysis is often used to check tests and surveys to see if all of the questions are correlated with the main issue being addressed or if all the questions along a claimed dimension are correlated to one another. If, as in the case of the FCI, a test was designed to measure an issue like student understanding of Newtonian force and the developers claimed the test measured student understanding across six dimensions as well as overall understanding of Newtonian Force, then one might expect that the students responses to correlate both within dimensions and overall. Huffman and Heller did a factor analysis of FCI data taken at the end of the mechanics section of introductory physics classes at the University of Minnesota and at a suburban high school in Minneapolis.

Their factor analysis yielded one factor that corresponded to one of Hestenes *et al's* Newtonian force dimensions (the Newton third law dimension) and part of a second (the kinds of force dimension). In general, they found that the student responses to the FCI in general were weakly correlated and therefore not more than loosely related. That is, the statistics of the student responses to the FCI did not correspond to Hestenes *et* *al.*'s taxonomy of the six conceptual dimensions shown in Table 4-1 that comprise the Newtonian force concept. Also, one would expect that since the FCI items all concern the Newtonian force concept, the student responses for the different items should correlate with one another. Because Huffman and Heller found that all of the items are only loosely related to one another, they were unable to statistically show that the force concept defined by Hestenes *et al.* is the same as the student force concept. In short, Huffman and Heller were unable to statistically verify that the students' responses showed a structure that was consistent with the defined structure of Newtonian force.

However, two things should be kept in mind. First, statistics are estimates and they are only one way to look at data. They are not the only way. Hestenes *et al*. interviewed students to verify not only that the students reasoning and responses were consistent in responding to the FCI items but also that the FCI result reflected the degree to which the students appeared to hold the Newtonian force concept as defined. Second, the weak correlations of the student responses are consistent with our understanding of how students structure their physics knowledge. These results come from physics education research on students' conceptual understanding and problem solving (see chapter 2).

The findings of Hestenses and Halloun, Minstrell,[42] and diSessa[43] among others indicate that student knowledge structures are often not coherent, well defined, consistent, or logically organized. The students' knowledge structure can be characterized as knowledge fragments where each fragment refers to a specific idea or situation. Both Minstrell's facets and diSessa's phenomenological primitives or p-prims can be characterized in this way.

In addition, student knowledge and reasoning is often context dependent. Almost every physics instructor has observed that many of their students fail to apply concepts and reasoning they have seen in class and homework assignments if they are given a problem in an unfamiliar context. In another example, physics education researchers have observed that students who appear to have acquired a Newtonian understanding of linear motion will revert to their common sense beliefs when first confronting rotational motion.[44] As we discussed in chapter 2, this implies that student knowledge does not exist in a common sense state or a Newtonian state, but rather that the common sense beliefs and the Newtonian concepts form a superposition state where the probabilities depend on what is triggered by the context of the situation or problem.

This is the heart of the 4-H controversy. The issue is not whether the Newtonian force taxonomy is a correct representation of the essential components of Newtonian force, but whether the taxonomy represents how students think about Newtonian force. The weak correlations between the items of the FCI found by Huffman and Heller suggest that if the FCI questions are measuring students' knowledge of the Newtonian force concept then "the FCI measures bits and pieces of students knowledge that do not necessarily form a coherent force concept."[45] However, it is still useful to use the Hestenes *et al*. taxonomy of the Newtonian Force concept and their taxonomy of misconceptions to see what aspects of Newtonian force students approach correctly and where and how the students use their common sense beliefs. One need only keep in mind that the two taxonomy structures may not represent how the students think about Newtonian force. (Please see the crystal axes analogy in chapter 5 on page 180.)

B. Steinberg & Sabella's Comparison of FCI Responses & Exam Problems (1997)

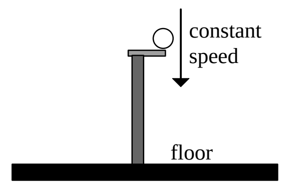
Noting that only a few studies have been published on how student performance on the FCI correlates with other measures of students conceptual understanding, Steinberg and Sabella recently reported on a comparison of student responses on the FCI and two qualitative exam problems.[46] The two qualitative exam problems are shown in Figure 4-3. The concepts needed to solve these problems correspond to concepts addressed by several questions on the FCI by design as shown in Table 4-3 below. The elevator problem was written by Steinberg; the two-cart problem was written by the author of this dissertation. The two problems were given on final exams in consecutive semesters one week after the students took the FCI. Because of the relatively small number of student responses, Steinberg and Sabella suggest that the results of their study may not be statistically significant. However, they feel that the results "give an indication of the types of issues that should be important to instructors using multiple-choice instruments."

In their analysis, Steinberg and Sabella looked at how individual students did the exam problems and the comparable FCI problems. They found that while there was a correlation between student performance on the exam and the comparable FCI questions, for certain students and certain questions the responses and the reasoning varied greatly. For example, both part A of the elevator problem and FCI question 18 ask about the forces on an elevator moving with constant velocity. Only 54% of the students correctly answered FCI question 18, although 90% of the students answered part A of the elevator problem correctly even though the situations are identical. The majority of incorrect responses indicated a net force in the direction of motion which

Figure 4-3a. Elevator problem that corresponds to several FCI questions from Steinberg and Sabella, “Performance on multiple-choice diagnostics and complementary exam problems,” in *The Physics Teacher*.[47]

Exam problem 1: Ignore all friction and air resistance in this problem.

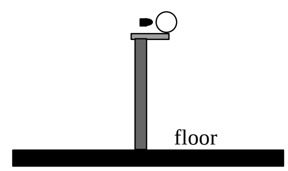
A. A steel ball resting on a small platform mounted to a hydraulic lift is being lowered at a constant speed, as shown in the figure at right.



i. Draw a free body diagram of the ball. Describe each type of force.

ii. Compare the magnitudes of the forces you have drawn. Explain your reasoning.

B. As the ball is moving down, a bullet moving horizontally hits the exact center of the ball (see figure at right) and then ricochets straight back. This causes the ball to immediately fall off the platform.



i. Draw a free body diagram of the ball after it is no longer in contact with the bullet or the platform. Describe each type of force.

ii. A vector that represents the velocity of the ball just before the bullet hits is shown below. Draw vectors that could represent the velocity at each of the 2 other times indicated. The scales of the 3 vectors should be consistent with each other. Explain your reasoning.

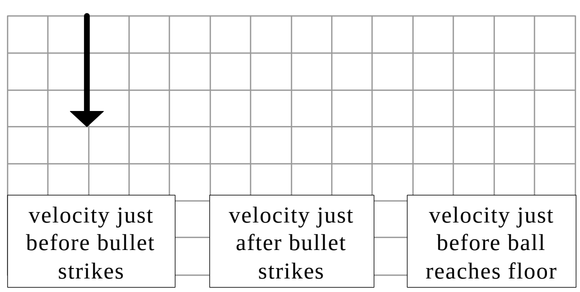
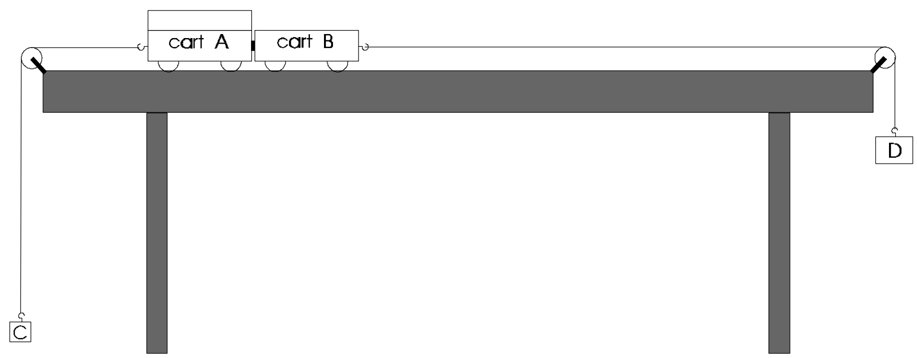


Figure 4-3b. Final exam problem that includes Newton's third law from Steinberg and Sabella, "Performance on multiple-choice diagnostics and complementary exam problems," in *The Physics Teacher*.[48] (Note: this problem was written by the author of this dissertation)

Exam Problem 2: Two carts, A and B ($Mass_A > Mass_B$), are placed on a table then stuck together with Velcro. Using pulleys, two small blocks, C and D ($mass_C < mass_D$), are connected by light strings to the carts as shown below. Initially, the carts are held in place. Ignore all friction in this problem.



At t = 0, the carts are released.

A. Draw and label two separate free-body diagrams, one for each cart, for a time after the carts start moving but before the Velcro pulls apart.

B. Rank all the horizontal forces from both your diagrams by magnitude, from largest to smallest. Explain the reasoning that you used to rank the forces.

Table 4-3. Comparison of concepts needed to solve exam problems and FCI items

Concept	Problem	Comparable FCI items
Newton's 1st Law	Elevator Problem – Part A	Question 18
Newton's 2nd Law	Elevator Problem – Part B(i)	Questions 9 & 22
Motion Diagrams	Elevator Problem – Part B(ii)	Questions12 & 28
Newton's 3rd Law	2 Cart Problem – Part B	Questions 2, 11, 13, & 14

is inconsistent with Newton's first law of motion. On the FCI questions and the part of the elevator problem that ask students to identify the forces acting on an object in motion after a brief impulse, only two thirds of the students answered the three virtually identical items consistently. The motion diagrams use different representations in the FCI questions and on the exam although the situations described are physically similar. While 43% of the students had responses that were consistent, another 43% wrote answers on the exam that corresponded to FCI choices that were different from the ones they selected.

On the FCI questions and the part of the exam problem in figure 4-3b that checks students' understanding of Newton's third law (N3), the student responses on the FCI questions and on the exam problem correlate, but not perfectly. Steinberg and Sabella note that only 65% of the students who answered three or four of the FCI items correctly were able to identify and compare the N3 force pair, 53% of those who answered two of the FCI items correctly, 29% of those who answered only one of the four FCI items correctly, and 14% of the students did not answer any of the FCI items correctly. Moreover, on FCI question 13, which describes a situation very similar to the exam problem where two objects remain in contact and accelerate uniformly for the entire motion, roughly 50% of the students answered the two consistently. Twenty-one of the students who answered the exam problem incorrectly clearly stated their reasoning. Of these, only six answered FCI question 13 in a way that was consistent with their reasoning on the exam.

Steinberg and Sabella offer the following explanations as to why the performance of a large minority of students does not seem to correlate:

1. Student models for physical systems can be ill formed and inconsistent.
2. The FCI looks colloquial, the exam formal.
3. The FCI is multiple-choice, the exam is open-ended.

All three of these arguments have merit. The first explanation recognizes that students often respond differently to problems that appear to physicists to be basically the same problem with different surface features. This is consistent with the earlier discussion of the structure of students' knowledge and the roles of cues for triggering different responses. The third explanation is simply recognition that students sometimes respond differently to questions where they have to come up with an answer and an explanation on their own as opposed to selecting from a set of choices. The choices can trigger responses that students might not come up with on their own.

The FCI questions are worded like real world experiences, the qualitative exam problems sound more like problems studied in a physics class. In the former, the students are asked a question about forces; in the latter, they are asked to draw free body diagrams and rank the forces. The wording and nature of the exam problems is more likely to trigger a classroom physics response. Another difference between the FCI and the exam questions in this study was the context in which they were taken by the students. The students took the FCI in the last week of classes and were told they would receive credit for participation but not for the correctness of their answers. In the final exam, the students know their grades depend on their demonstrating what they learned in class. I suggest that this makes it much more likely for students to use their formal physics knowledge on the exam rather than their common sense beliefs. I will discuss this point in chapter 9.

VII. THE FORCE AND MOTION CONCEPT EVALUATION (FMCE)

The FCI gives a good indication of a student's overall understanding of the Newtonian concept of force and it can be used to gauge common-sense beliefs held by the class as a whole. However, because some issues are addressed by only a few questions in one or two contexts, it is not a precise enough instrument to determine which common-sense beliefs are held by a single student or under what circumstances they are used. In this sense the FCI is a survey instrument. The results of a survey instrument can indicate possible problems but lack the resolution to be a reliable measure of a single student. A diagnostic instrument may address fewer issues than a survey instrument but covers each issue thoroughly in multiple contexts.

To aid their development of MBL laboratory and demonstration curricula,[49] Thornton and Sokoloff wanted an instrument that would give more information about the common sense beliefs of each student than was possible with the FCI, i.e. a diagnostic test of Newtonian concepts. They developed the Force and Motion Conceptual Evaluation (FMCE) to address this need (a copy of the FMCE can be found in Appendix B).[50] The FMCE does not cover as much material as the FCI (for example, there are no questions on circular motion) but uses more questions for each concept and approaches them in a few different contexts. Also, the FMCE places more emphasis on students' understanding of graphical representations of velocity, acceleration, and force. A few of the schools participating in this study used the FMCE instead of the FCI to measure students' conceptual understanding of mechanics.

A few of the FMCE questions are identical to items on the FCI; the remainder were developed through student interviews and testing open-ended versions of the questions with explanations. Like the FCI, the FMCE distractors come from the most common student responses to open-ended versions of the questions. To test the validity, Thornton and Sokoloff have given the FMCE to hundreds of physics faculty, compared student responses on multiple-choice versions and open-ended with explanation versions of the FMCE, and asked additional questions on exams to compare with the FMCE results. The faculty agree with the interpretation and Newtonian response to all 43 questions.[51] There is an extremely high correlation between the student responses to the various styles of questions, particularly the multiple-choice and open-ended with explanation versions of the FMCE questions (>90%). In addition, the pre and post results have been very stable and repeatable, comparing equivalent classes at several different schools for both traditional and enhanced instruction.

Thornton and Sokoloff have reported FMCE results that are very similar to the FCI results discussed earlier in the chapter, i.e. that traditional classes show little gain (less than 10% increase in the total score) while classes using research-based curricula show much greater gains. Thornton and Sokoloff note that patterns of individual student responses are usually consistent with either a Newtonian understanding or specific common-sense beliefs about motion. Their results also show that many students do not apply a consistent model to questions on force and motion. In their words, "many students view speeding up, slowing down, moving at constant speed, and standing still to be independent states of motion that do not require a consistent relationship between force, acceleration, and velocity."

Since a few schools only used the FMCE to measure conceptual understanding, one may ask whether the FMCE and the FCI are comparable. In the 1995 spring semester at University of Maryland, the students in the first semester of the introductory physics for engineering majors sequence took the FCI on the first day of class and the FMCE in recitation sections later in the same week. I found the overall scores have a correlation (Pearson product coefficient) of 0.79 which is high enough for a direct comparison of the two scores.

In chapter 9, the results from the velocity graph and the Newton 3 questions from the FMCE will be discussed in detail in addition to the overall FMCE results.

VIII. WHAT DO MULTIPLE CHOICE CONCEPT TESTS TELL US ABOUT STUDENTS' CONCEPTUAL UNDERSTANDING?

The discussion in this chapter indicates that tests like the FCI appear to measure students' understanding of basic concepts compared with their common-sense beliefs. Both the FCI and FMCE have proved their validity and reliability through extensive testing and comparisons with interviews and open-ended versions of the questions. In addition the results have been shown to be very robust in classes at many institutions. The results of pre- and post-course testing indicate that classes taught with traditional lecture instruction show small gains on these tests, while classes using research based active-engagement activities resulted in significantly better gains. Validation studies have shown that the results are consistent with students' understanding of basic Newtonian concepts; although factor analysis suggests that student knowledge is fragmented. Students' conceptual knowledge can appear to be a mixture of both Newtonian and common-sense beliefs.

In addition, while the FCI results were strongly correlated with the exam problems in the Steinberg and Sabella study,[52] the responses for certain students and certain problems differed greatly. This implies that tests like the FCI may not necessarily be a good measure of students' ability to use concepts in problems, i.e. test the functionality of their conceptual knowledge. Furthermore, student responses to questions that test their understanding of concepts appear to be triggered by cues in the context of the problem in question and the test itself. More research is needed to see how the FCI compares with measures of students thinking and reasoning using physics concepts.

The FCI has played an important role of convincing the physics education community of the extent of student difficulties with conceptual understanding of Newtonian force and limited effect of traditional lecture instruction. However, while the FCI is a useful measure of students' conceptual understanding, there is a tendency in the community to rely solely on tests like the FCI. This over-simplifies the view of both student learning and assessment. The FCI should be one part of a broad-based comprehensive assessment.

---

[1] D.E. Trowbridge and L.C. McDermott, "Investigation of student understanding of the concept of velocity in one dimension," *Am. J. Phys.* **48**, 1020-1028 (1980); D.E. Trowbridge and L.C. McDermott, "Investigation of student understanding of the concept of acceleration in one dimension," *Am. J. Phys.* **49**, 242-253 (1981); L.C. McDermott, "Research on conceptual understanding in mechanics," *Phys. Today* **37** (7), 24 (July 1984).

[2] L. Viennot, "Spontaneous reasoning in elementary dynamics," *Eur. J. Sci. Educ*. **1**, 205-221 (1979).

[3] I. Halloun and D. Hestenes, "Common sense concepts about motion," *Am. J. Phys.* **53** (11), 1056-1065 (1985).

[4] J. Clement, "Students' Preconceptions in Introductory Mechanics," *Am. J. Phys.* **50** (1)**,** 66-71 (1982).

[5] I. Halloun and D. Hestenes, "The initial knowledge state of college physics students", *Am. J. Phys.* **53** (11), 1043-1055 (1985).

[6] D. Hestenes, M. Wells, and G. Swackhamer, "Force Concept Inventory," *Phys. Teach*. **30** (3), 141-158 (1992).

[7] Some examples are the *Heat and Temperature* concept test and the *Circuits* concept test written by Thornton and Sokoloff (unpublished, but available from the authors) as well as the Wave Concept Test being developed by Michael Wittmann at University of Maryland. There are efforts currently underway by Van Heuvelen, Heiggelke, Maloney, and Beichner to write a concept test for electricity and/or magnetism.

[8] R.K. Thornton and D.R. Sokoloff, "Assessing student learning of Newton's Laws: The Force and Motion Evaluation and the evaluation of active learning laboratory and lecture curricula," *Am. J. Phys*. **66** (4), 338-351 (1998).

[9] D. Huffman and P. Heller, "What does the Force Concept Inventory actually measure?" *Phys. Teach*. **33** (3), 138-143 (1995).

[10] R.N. Steinberg and M.S. Sabella, "Performance on multiple-choice diagnostics and complementary exam problems," *Phys. Teach*. **35** (3), 150-155 (1997).

[11] The 4 H's are Halloun, Hestenes, Heller, and Huffman. The controversy name comes from two articles on this topic published in the same issue of *The Physics Teacher*. One article was a response by Halloun and Hestenes to an earlier article by Heller and Huffman and the other article was Heller and Huffman's counter-response.

[12] See Ref. 9.

[13] I. Halloun and D. Hestenes, "Interpreting the Force Concept Inventory: A response to the March 1995 critique by Huffman and Heller ," *Phys. Teach*. **33** (8), 502-506 (1995).

[14] P. Heller and D. Huffman, "Interpreting the Force Concept Inventory: A response to Hestenes and Halloun ," *Phys. Teach*. **33** (8), 503-511 (1995).

[15] F. Reif, "Millikan Lecture 1994: Understanding and teaching important scientific thought processes," *Am. J. Phys*. **63** (1), 17-32 (1995).

[16] Private communications with Pat Heller at University of Minnesota and Edward Redish at University of Maryland.

[17] See Ref. 6.

[18] See Ref. 6.

[19] The correct Newtonian response to item 12 (d) requires an understanding of buoyancy that escapes even some physics graduate students. Because of this difficulty and since buoyancy is not even covered in some introductory classes, response (b) is also considered a Newtonian response.

[20] D.J. Mueller, *Measuring Social Attitudes: A Handbook for Researchers and Practitioners* (Teachers College, Columbia University, New York, 1986).

[21] G.F. Kuder and M.W. Richardson, "The theory of the estimation of test reliability," *Psychmetrika* **2**, 151-160 (1937).

[22] See Ref. 3.

[23] See Ref. 6.

[24] See Ref. 13.

[25] D. Hestenes and M. Wells, "A mechanics baseline test," *Phys. Teach*. **30** (3), 159-166 (1992).

[26] See Ref. 13.

[27] R.K. Thornton, "Using large-scale classroom research to study student conceptual learning in mechanics and to develop new approaches to learning," in *Microcomputer-Based Labs: Educational Research and Standards*, edited by R.F. Tinker, *Series F, Computer and Systems Sciences* **156** (Springer Verlag, Berlin, Heidelberg, 1996), 89-114.

[28] See Ref. 5.

[29] Variance = (Pearson's correleation coefficient r)^2

[30] See Ref. 5.

[31] See Ref. 6.

[32] R.R. Hake, "Active-engagement vs. traditional methods: A six thousand student study of mechanics test data for introductory physics courses," *Am. J. Phys*. **66** (1), 64-74 (1998).

[33] R.R. Hake, "Towards mastery of Newtonian mechanics by the average student," AAPT Announcer **24** (1), 23 (1994, abstract only).

[34] E.F. Redish, J.M. Saul, and R.N. Steinberg, "On the effectiveness of active-engagement microcomputer-based laboratories," *Am. J. Phys.* **65** (1), 45-54 (1997).

[35] D.C. Howell, *Statistical Methods for Psychology*, 3rd Ed. (Duxbury Press, Belmont CA, 1992) 181-185.

[36] See Ref. 33.

[37] R.R. Hake, "Active engagement methods in introductory mechanics courses," preprint, April 1997, available upon request.

[38] Phys-L is a listserve for undergraduate and high school physics instructors to discuss issues related to physics teaching. PhysLrnR is a listserve for discussing experimental and theoretical physics education research topics.

[39] See Ref. 31.

[40] Private communication with R. Hake, October 1997.

[41] See Refs. 9, 11, 13 & 14.

[42] J. Minstrell, "Facets of students' knowledge and relevant instruction," in *Research in Physics Learning: Theoretical Issues and Empirical Studies*, edited by R. Duit, F. Goldberg, and H. Niedderer (University of Bremen, Germany, 1991), 110-128.

[43] A. diSessa, *Cognit. Instruct*. **10** (2 and 3), 105-225 (1993)..

[44] G. Francis and M. Larson, "Student misconceptions in rotational kinematics and dynamics," *AAPT Announcer* **24** (2), 64 (July 1994, abstract only); G. Francis, "Spiraling back with rotational mechanics: transferring conceptual change from linear to rotational motion," *AAPT Announcer* **25** (4), 44 (December 1995, abstract only).

[45] See Ref. 14.

[46] See Ref. 10.

[47] See Ref. 46.

[48] See Ref. 46.

[49] D.R. Sokoloff and R.K. Thornton, "Using interactive lecture demonstrations to create an active learning environment," *Phys. Teach*. **35** (6), 340-347 (1997); D.R. Sokoloff, R.K. Thornton, and P.W. Laws, *RealTime Physics Mechanics* V. 1.40 (Vernier Software, Portland OR, 1994); R.K. Thornton and D.R. Sokoloff, *Tools*

for Scientific Thinking - Motion and Force Curriculum and Teachers' Guide, 2nd ed. (Vernier Software, Portland OR, 1992).

[50] See Ref. 8.

[51] Up to forty percent of physics faculty initially fail to pick the correct answer for question #6, although they change their response to the correct answer when challenged. Thornton and Sokoloff have been unable to come up with a new wording for this problem and suggest that question 6 be evaluated in conjunction with other questions on force and acceleration.

[52] See Ref. 44.