

# Interpreting the Force Concept Inventory

## A Response to March 1995 Critique by Huffman and Heller

By David Hestenes and Ibrahim Halloun\*

Department of Physics & Astronomy, Arizona State University, Tempe, AZ 85287-1504

The Force Concept Inventory (FCI)<sup>1</sup> is a unique kind of "test" designed to assess student understanding of the *most basic* concepts in Newtonian physics. It can be used for several different purposes, but the most important one is to evaluate the effectiveness of instruction. For that purpose, the FCI is probably the most widely used instrument in physics education today. Results of many independent investigations have been reported at the biannual AAPT meetings since the FCI was published in March 1992. Including unpublished data that have been brought to our attention, we estimate that the FCI has been administered in classes of well over a hundred different teachers to more than ten thousand students in high schools, colleges, and universities. For comparative analysis, Richard Hake<sup>2</sup> has been collecting data on university and high-school physics taught by many different teachers and methods. Readers with a similar concern are urged to contact him.

Douglas Huffman and Patricia Heller (H&H)<sup>3</sup> recently published a "factor analysis" of FCI data and have claimed that it raises serious concerns about the validity and interpretation of the FCI. We find that their data provide support for our position, but their concerns are unjustified. Moreover, they have overlooked relevant analysis of the issues in our published papers, and we take issue with their advice on using the FCI. In the following we take familiarity with Refs. 1 and 3 for granted; readers who have not studied them are encouraged to do so.

### Design and Development of the FCI

The FCI has a predecessor, the Mechanics Diagnostic Test (MDT),<sup>4</sup> which has also been widely applied. About 60% of the FCI is the same as the MDT, and the results from both tests are perfectly consistent and mutually supportive. Analysis of MDT results led to the improvements in the FCI. Accordingly, we regard the FCI as an improved version of the MDT rather than a completely new test. We mention this because the data and analysis in our two papers on MDT<sup>4,5</sup> have strong bearing on the interpretation of the FCI and its results. For a full understanding of what has gone into the development of the FCI, these papers should be consulted.

The main advance in the FCI over the MDT comes from a more systematic analysis of the basic concepts in introductory Newtonian mechanics. The FCI covers those concepts

more comprehensively and facilitates interpretation of the results. The conceptual analysis is outlined in Tables I and II of Ref. 1, which the serious reader is invited to review. The two tables are complementary. Table I outlines the Newtonian concepts involved and notes the test items where each one is involved. Table II supplies a taxonomy of non-Newtonian responses to the test questions. These tables are crucial to interpreting results of the FCI.

Since H&H question the relevance of Table I and they ignore Table II, it is clear that the whole matter of interpretation must be carefully reconsidered. *Their own data is actually in complete accord with our results and our interpretation of the FCI.* We welcome their data especially because it strongly supports a very important point about which our advice has often been overlooked or ignored, namely, that, for best results, the FCI should be administered and interpreted as a whole; separate pieces of it are much less reliable and informative. The reasons for this advice are worth reviewing.

### Validity of the FCI

The concept of force is central to Newtonian physics. It is also very complex. Table I of Ref. 1 analyzes the *Newtonian force concept* into six conceptual dimensions, each of which has additional structure. We maintain that all six dimensions in Table I are *essential* to the Newtonian force concept, and the FCI was designed to probe each of them. With one exception, to every question on the FCI there is precisely one and only one Newtonian response among the five alternatives. Since the test was published it has been carefully examined by many physics professors. Suggestions have been made to improve the wording or diagrams for a few of the questions,<sup>2</sup> but there has been no serious question as to which is closest to a Newtonian choice on any of them. The *face validity* of the test is thus beyond reasonable doubt.

Having built the FCI to survey the whole range of concepts in Table I of Ref. 1, we are justified in interpreting the total score of Newtonian responses as a measure of the degree to which the student has assimilated the Newtonian force concept. To put it more succinctly, *the FCI score is a measure of one's understanding of the Newtonian force concept.*

Continued on page 504.

How accurate is this measure? That is a question about content validity of the test. To answer it we must estimate the probability of false negatives and false positives.

The answer to a given question is said to be a *false negative* if a Newtonian thinker has chosen a non-Newtonian response. An answer is a *false positive* if a Newtonian response has been chosen for non-Newtonian reasons. A major problem in multiple-choice test development is to minimize false positives and negatives. In that, the FCI has been exceptionally successful because of special features in its design.

From our qualitative analysis of responses by Newtonian thinkers, we judge the probability of a false negative to be certainly less than ten percent (fewer than three questions “missed”). This is a very conservative estimate. The Newtonian response to most questions is so obvious and unproblematic to Newtonian thinkers that false negatives can only be attributed to carelessness or inattention. We have confirmed that with many interviews. To support this conclusion with statistical data on the scores of physics professors would be overkill.

The minimization of false positives is more difficult. Obviously they cannot be eliminated altogether—even random choices have a 20% chance of false positives. Student choices are not random, however, as is clear from interviewing them—they usually have definite reasons for their choices. In the FCI design we used two devices to reduce the “noise” from false positives. First, the FCI probes each conceptual dimension with several questions involving different contexts and viewpoints. A false positive on one of the questions can then be partially compensated for by a non-Newtonian choice on another. Second, we introduced *powerful distracters* onto each FCI question, namely, non-Newtonian alternatives that appear eminently reasonable to students, because they were culled from extensive student interviews.<sup>1,4,5</sup>

In the light of all this, it is surprising to find H&H concluding from factor analysis data that “The items on the inventory appear to be only loosely related to each other, and instructors should be cautious about concluding that the inventory actually measures students’ understanding of a ‘force concept’. ...The fact that the items did not group together on the six conceptual dimensions of the force concept indicates that the FCI should not be decomposed into the six dimensions originally proposed by its authors.”

On the contrary, we see their data as completely consistent with our own and supportive of our interpretation. In the first place, however, their data is irrelevant to the question of whether the FCI score is a valid measure of the Newtonian force concept, because it was clearly gathered from a predominantly non-Newtonian population. To address this question statistically they would need a certified Newtonian population, such as a group of physics professors. In that case, we guarantee that they would find near perfect clustering of their

data about a single factor, every question correlating almost perfectly with every other one.

In the second place their data has no bearing on the purely logical decomposition of the FCI into six conceptual dimensions. The fact that their data do not cluster on the six dimensions simply means that any “Newtonian signal” they might contain is swamped by the noise of false positives. To extract a signal from the noise, they could, for example, perform a factor analysis on the group of students with total FCI scores between 60 and 80%, and separately for the group with scores greater than 80%. What they might expect to find will be indicated later.

### *Student Concepts of Force*

The H&H data show clearly that student responses do not cluster on the six conceptual dimensions of Table I.<sup>1</sup> They rightly conclude that Table I does not describe the structure and organization of *student* force concepts. However, this is no reason to question the value of Table I for interpreting FCI data. Contrary to what H&H suggest, Table I was never intended to describe student concepts. Rather it describes the Newtonian standard against which student concepts can be *compared in detail*. It is precisely this that justifies interpreting the FCI score as a measure of the *disparity between student concepts and the Newtonian force concept*.

While Table I describes the Newtonian force concept, Table II<sup>1</sup> classifies alternative student concepts. To facilitate comparison with the Newtonian concept, Table II is organized into six categories corresponding to those in Table I. We do not claim, however, that these categories describe conceptual structures of individual students. That should be clear from our extensive discussion of student concepts.<sup>1,5</sup> Table II lists a heterogeneous set of student concepts in each category, some of which are mutually contradictory. Moreover, we have discussed a number of specific examples where student concepts cut across and conflict with Newtonian categories.<sup>1,5</sup> We need not repeat the details, but it seems necessary to mention these facts because H&H ignore Table II completely and go on to discuss student concepts as if we had never said a word on the subject.

In describing the well-established findings of educational research that *student beliefs about physics are loosely organized, incoherent, ill-defined and context-dependent*, H&H have failed to note that we have taken pains to incorporate these insights into the design and interpretation of the FCI. Indeed, from our own extensive analysis of test data and student interviews we have concluded that: (1) Student concepts are often “*vague and undifferentiated*,”<sup>5</sup> and they are “*incompatible with Newtonian concepts in most respects*.”<sup>1</sup> (2) Student belief systems are *incoherent*<sup>1,5</sup> and “*can best be described as bundles of loosely related and sometimes inconsistent concepts*.”<sup>5</sup>

Naturally, we all want deeper insight into the cognitive structure of student beliefs. H&H mention the important work of diSessa on this issue. They fail to note that our account of metaphors in student reasoning is nearly equivalent to diSessa’s “*phenomenological primitives*,” because the

meanings of metaphors are rooted in personal experience. We have identified prominent roles for three specific metaphors in student reasoning about forces.<sup>1</sup> This is not the place to repeat or extend the analysis. We mention it because we believe it is a key to understanding how students think.

The FCI data has a lot of information about student concepts that might well be extractable and clarified by cogent statistical analysis. But that cannot be done by the H&H approach, because apparently they have only analyzed correlations among Newtonian responses. There is little information about non-Newtonian concepts in that. To find out what the FCI has to say about the alternative concepts of students, the non-Newtonian responses classified in Table II must be studied. We have done this with our own data by qualitative means, and some of our conclusions are reviewed in the next section. We would welcome a more rigorous statistical analysis, but that would be difficult, because the response pattern of each student must be analyzed separately and compared to identify possible groups of students with similar patterns. However, FCI data is exceptionally rich, robust, and informative, so the results might be worth the effort.

### **What Does the FCI Score Tell Us?**

From a physics perspective, each FCI question requires the student to discriminate a Newtonian answer from four alternative non-Newtonian responses. The answers appear to be so elementary and obvious to physicists that, on a first pass, few would consider the questions worth asking. This makes a *negative* (non-Newtonian) response highly informative. In shock, many a physics teacher has exclaimed "How could *my* students miss *that*?" For introductory physics courses, the FCI scores are invariably much lower than the instructor expects. The *surprise value* increases with each question to make the total score highly informative, the more so because the FCI probes the whole range of the most basic concepts. Extensive interviews of students by many investigators have repeatedly confirmed that a negative response is nearly always a reliable indicator of some deficiency in the student's understanding of Newtonian concepts.

A positive (Newtonian) response to a single question is, of course, much less informative than a negative one. The likelihood that it is a false positive decreases, however, as the number of related positives increases. A near perfect score is therefore a strong indicator of Newtonian thinking. It is not a perfect indicator, of course, because Newtonian physics requires more than recognition skills. There is strong evidence, however, that the FCI score is highly correlated with other "Newtonian skills," such as problem solving.<sup>6</sup> On the basis of such data we interpret an FCI score of 85% as the *Newtonian Mastery threshold*. We are confident in identifying students with scores above this threshold as *confirmed Newtonian thinkers*.

We suggest an FCI score of 60% as the *entry threshold* to Newtonian physics. Students who have just reached this threshold have barely begun to use Newtonian concepts coherently in their reasoning. From our test results and stu-

dent interviews, we can describe the thinking of students below this threshold in terms of the following typical characteristics: (1) *undifferentiated concepts* of velocity and acceleration; lacking a vectorial concept of velocity; (2) lacking a *universal force* concept (i.e., believing that there are other influences on motion besides forces), and unable to reliably identify the agents of forces on an object; (3) *fragmented and incoherent concepts* about force and motion.

The above interpretation of FCI scores is consistent with a *three-stage model of conceptual evaluation* in learning Newtonian mechanics. In Stage I students develop a *universal force* concept and learn to identify active and passive agents of force. Completion of this stage is roughly indicated by an FCI score of 60%. In Stage II students develop *coherent dynamical concepts*, including vectorial concepts of velocity, acceleration, and force. In Stage III students develop a *complete interaction concept*. This involves a full understanding of Newton's Third Law.

Of course, the conceptual development of individual students is influenced by the order in which concepts are introduced in instruction. But our data suggests that there is a natural order in which concepts are most easily learned. This is a worthy issue for more educational research and for experimentation by teachers.

### **Using the Force Concept Inventory**

We have discussed the various uses for the FCI at some length before.<sup>1,4</sup> However, from our observations of uses by others over recent years, it appears that some points need more emphasis. The most important point is this: *For the purpose of course and teaching evaluation, the entire FCI test should be used.* There are two good reasons why. First, the total FCI score has proved to be a useful measure for comparing different courses and teaching methods, and a large database will therefore facilitate comparisons throughout the teaching community. Second, as explained above, the total FCI score is the most reliable single index of student understanding, because it measures coherence across all dimensions of the Newtonian force concept. The H&H factor analysis data supports this conclusion by documenting the incoherence in non-Newtonian responses.

The distribution of individual FCI scores should be considered, not just the mean score for the whole class. If the FCI is administered both as a pretest and a posttest, the *FCI gain* can be computed for each student. This is an especially informative number if it is either large [ $> 60\%$  of the maximum possible gain ( $100 - \text{pretest score}$ )] or small [ $< 20\%$  of the maximum possible gain]. Average gains under conventional instruction are consistently close to 25% of the maximum possible gain. The class distribution of gains may show differences in the effect of instruction on "weaker" vs "stronger" students.

Concerning the use of the FCI as a placement exam, we feel that our views have been misrepresented by H&H. The purpose of a placement exam is to predict performance so students likely to have undue difficulty can be identified. On the basis of our published evidence,<sup>4</sup> we are confident in

asserting that the FCI, coupled with a simple math test, is probably the most accurate available predictor of performance in introductory physics in either high school or college. Nevertheless, we advise against using it as a placement exam for such courses, because its high predictive power is indicative of *inadequacies in the instruction rather than in the students*.

However, as a placement exam for accelerated or advanced courses the FCI may be very useful. We expect students testing below the Newtonian entry threshold (60%) to have difficulty with such courses. Our limited data supports this expectation. We have used the FCI several times as a pretest for an Honors section of university physics, wherein all the students have exceptional academic records. There is wide distribution in their FCI scores, however. Without exception, students testing below the 60% threshold have difficulty with the course while those above did not. We also have evidence that below-threshold high-school students do not fare well in textbook-based Advanced Placement courses, but that deserves more study.

In collaborative settings, an FCI pretest (coupled with a mathematics pretest) can also be useful for putting teams of students together. Based on the two pretests, students can be classified into three (or four) competence levels.<sup>4</sup> Teams can

then be formed with a distribution of competence levels. We have found that such heterogeneous teams work successfully, but again this deserves more study.

*\*On leave from Lebanese University.*

#### References

1. D. Hestenes, M. Wells, and G. Swackhamer, "Force concept inventory," *Phys. Teach.* **30**, 141–151 (1992).
2. Richard Hake, "Survey of test data for introductory mechanics courses," *AAPT Announcer* **24** (2), 55 (July 1994). Halloun, Hake, and Mosca have recently produced an improved version of the FCI. However, the revision does not significantly affect numerical results.
3. D. Huffman and P. Heller, "What does the force concept inventory actually measure?" *Phys. Teach.* **33**, 138–143 (1995).
4. I. Halloun and D. Hestenes, "The initial knowledge state of college physics students," *Am. J. Phys.* **53** (11), 1043–1055 (1985).
5. I. Halloun and D. Hestenes, "Common sense concepts about motion," *Am. J. Phys.* **53** (11), 1056–1065 (1985).
6. D. Hestenes and M. Wells, "A mechanics baseline test," *Phys. Teach.* **30**, 159–166 (1992).
7. Stephen Jay Gould, *The Mismeasure of Man* (Norton, New York, 1981).

### Appendix: What Does Factor Analysis Actually Measure?

We have explained why the Heller and Huffman factor analysis results actually strengthen our interpretation of FCI data, but we feel that a word of caution about the use of factor analysis is in order. We believe that the uncritical account of factor analysis by H&H seriously overestimates its power. Thus, they assert that "By examining the items that group together on each factor, one can determine if a test actually measures the concepts it appears to measure." If only it were so easy, it could have saved us a lot of trouble in validating the FCI. In actual practice, however, the qualifications are so severe that their assertion loses most of its strength. An incisive critique of factor analysis (and other statistical techniques) is given by Stephen Gould in his very readable paperback.<sup>7</sup> It is, of course, important for teachers to be critically aware of the uses and abuses of statistics.

We do not have to consult the experts to recognize severe limitations of factor analysis. It is sufficient to examine the account by H&H in the appendix to their paper. We note there that the test items are assumed to be *linearly related* to a set of *uncorrelated* (hence *independent*) factors. Now, it is a mathematical theorem that such factors exist if the linear relation is assumed, but there is no guarantee whatsoever that these factors can be given a sensible interpretation. The linear relation is, of course, assumed for reasons of mathematical simplicity, and it can be justified only in very special cases. It certainly does not express the complex relations among concepts in the FCI. It is even questionable in the oversimplified example of H&H. They treat velocity and acceleration as simple independent variables. But they are certainly not conceptually independent, for there is no concept of acceleration without a concept of velocity. And what could be the meaning of their example factor analysis of unspecified velocity and acceleration questions? Are we to believe that the original velocity questions actually had a little acceleration mixed in which had to be separated out to get a "pure velocity factor?" In truth, "acceleration" is one of the most subtle and difficult concepts in introductory physics. We doubt that factor analysis has anything to contribute to evaluating questions that test for it.