Interpreting the Force Concept Inventory

A Reply to Hestenes and Halloun

By Patricia Heller and Douglas Huffman

Department of Curriculum and Instruction, University of Minnesota, Minneapolis, MN 55455-1208

Although the Force Concept Inventory (FCI) is a widely used instrument in physics education, in our article, “What Does the Force Concept Inventory Actually Measure?” we determined that the FCI does not meet one of the common standards of test construction. As viewed by expert Newtonian thinkers, the FCI appears to test students’ understanding of the Newtonian force concept, which is measured by questions dealing with different dimensions of Newtonian physics, such as Newton’s First Law, Second Law, Third Law, and so on. However, a standard correlational analysis (factor analysis) of students’ responses indicates that, from the students’ point of view, the FCI does not appear to test for a coherent, universal force concept, dimensions of a force concept, or any organized alternative beliefs (e.g., impetus). Based on our results, we concluded that although the FCI could be useful as a diagnostic tool and for course evaluations, caution should be used in interpreting the test scores and using the scores to make decisions about individual students.

In their response to our article, David Hestenes and Ibrahim Halloun (H&H) claim that our concerns about the validity and interpretation of the FCI are unjustified because (a) we overlooked relevant analysis of these issues in their published articles, and (b) factor analysis is an inappropriate statistical tool to draw conclusions about either the validity of the FCI or student concepts of force. They assert that our advice about caution in interpreting and using the FCI is ill considered. Not withstanding their objections to our analysis technique, they claim that our data support their position that the FCI measures coherence across all dimensions of the Newtonian force concept. Although we carefully studied all of the authors’ previous publications, we find that they have presented no data or analysis that contradicts either our analysis or our advice of caution in interpreting and using FCI scores. In the following we present our reply to each of these issues.

**Factor Analysis and Validity of the FCI**

Factor analysis is a statistical technique, based on the correlations between test items, that is widely used in the process of constructing and validating tests. Ideally, factor analysis is used early in the test development process to provide empirical evidence that can be used, along with theoretical criteria, to guide the selection of test items that appear to measure a particular theoretical construct. For example, to construct a test designed to measure students’ understanding of Newton’s Third Law, the test constructor could write numerous items that have strong face and content validity, pilot test the items with students, and then use factor analysis along with other item analysis techniques to guide the selection of items that appear to be the best measures of Newton’s Third Law. Although factor analysis is best used during the test development process, it can still be used after the fact, as we did in our article, to help provide information about how the test items are related from the students’ point of view. That is, since we use the FCI as one indicator in our efforts to improve introductory physics courses, and the authors of the FCI had not reported a factor analysis to investigate whether the theoretical dimensions on the FCI are supported in practice, we thought it necessary to conduct a post hoc factor analysis of the FCI scores.

Since H&H first published their articles about the Mechanics Diagnostic Test, we have been using this test and its successor, the FCI, to help us evaluate the effectiveness of our introductory physics courses at the University of Minnesota. Based on our reading of these four articles, we agreed with the authors that the FCI has face validity, and that the six conceptual dimensions, summarized on Table I of Ref. 2, are essential to the Newtonian force concept. We also acknowledged that the test has content validity; the authors did an outstanding job in selecting common-sense distractors, culled from student interviews conducted by themselves and other researchers, that minimize the possibility of false positives (a Newtonian response chosen for a non-Newtonian reason) and false negatives (a non-Newtonian response chosen by a “Newtonian” thinker). These distractors are summarized by the authors in Table II of Ref. 2.

We did not, in fact, either question the face validity of Table I or ignore Table II. Instead, our analysis addressed the issue of what the FCI items measure for the audience for which it is intended—students in high school and introductory college physics courses. This issue can only be addressed

Continued on page 507.
by analyzing the response patterns of the intended audience of students, and not, as H&H suggest, by analyzing the scores of a "certified Newtonian population such as a group of physics professors." Although it may be true that physics professors would get near perfect scores on the FCI (face validity), this information reveals little about what the test items measure for the intended audience of students. In other words, we agreed that the "theory" (face and content validity) of the FCI was reasonable. We used factor analysis as one method, which is commonly used in educational measurement, to test the "theory" on the intended audience. We were very surprised by the results.

**Factor Analysis and Student Concepts of Force**

H&H claim that we cannot extract information about students' alternative beliefs by a factor analysis because this approach "only analyzed correlations among Newtonian responses," and there is "little information about non-Newtonian concepts in that." On the contrary, a correlation between two items, by definition, takes into account both Newtonian (correct) responses and non-Newtonian (incorrect) responses of each student. To illustrate with a simple example how a correlation analysis can yield information about students' conceptual understanding, and conversely, how an average score can be misleading, consider a hypothetical two-item test designed to measure one aspect of students' understanding of Newton's First Law—that when no net force acts on an object, it moves in a constant straight line. For concreteness, imagine that the two items are #6 and #26 from the FCI. Item #6 requires students to identify the trajectory of a hockey puck on a frictionless surface after it receives an instantaneous kick. Item #26 requires students to identify the trajectory of a rocket after it has received a thrust, and the engines are turned off.

The premise of a correlation analysis is the following: If the items measure the same concept for students (in this case one aspect of Newton's First Law), then the students' responses to these items should be correlated. That is, students who understand this aspect of Newton's First Law would tend to consistently select the powerful but incorrect non-Newtonian distracters. Consider two hypothetical samples of students, one in which 40% of the students answered correctly, and one in which 80% of the students answered correctly. For discontinuous data, the correlation coefficients can be calculated from a table of cell frequencies, as illustrated in Figs. 1 and 2:

$$r = \frac{ad - bc}{\sqrt{(a + b)(c + d)(a + c)(b + d)}}$$  

(1)

where $a, b, c,$ and $d$ are cell frequencies.
Figures 1a and 2a show, for the two hypothetical samples of students, the cell frequencies for a perfect correlation ($r = 1.00$): all students answer both items correctly or both items incorrectly. Regardless of whether the average score on the two-item test is low (40%) or high (89%), a perfect correlation would support (but not prove) the interpretation that the two items measure the same concept from the students’ point of view, namely that when no forces are acting on an object, it moves in a constant straight line. Of course, a perfect correlation is rarely achieved in practice because of false negatives (incorrect responses chosen by students who do understand Newton’s First Law) and false positives (correct responses chosen by students who do not understand the First Law). For this hypothetical, two-item test, assume that there are no false negatives, and a 10% chance of a false positive. We will call the resulting correlations for this case “strong” correlations, as shown in Figs. 1b and 2b. These results would also support (but not prove) an interpretation that the items measure the same concept for the students. The average score on the hypothetical test (40% and 80% for each sample) reflects the approximate percentage of students who answered both items correctly, so we could infer that the test average tells us the approximate percentage of students who understand one aspect of Newton’s First Law. Moreover, since most of the students who answered incorrectly answered both items incorrectly, we might infer that these students have a coherent alternative belief, perhaps the “impetus” theory from Table II (in Ref. 2) or the belief that the last force to act on an object determines its motion.

This interpretation of the average score is valid only if there is a strong correlation between the items. A “fairly strong” correlation, as shown in Figs. 1c and 2c, is more difficult to interpret, since it could arise for several reasons. One interpretation is that the items are not very well constructed, so there are both more false positives and more false negatives than intended. Another interpretation is that the students who understand Newton’s First Law answer both items correctly, but the distracters are not capturing students’ alternative beliefs—the students who do not understand Newton’s First Law are answering the questions essentially randomly. (This interpretation was used to calculate these “fairly strong” correlations.) Nevertheless, even this case would support (but not prove) an interpretation that the test items are fairly well able to measure students’ understanding of a First Law concept. More care would be needed, however, in interpreting the total score on the hypothetical two-item test. The percentage of students who understand Newton’s First Law is considerably smaller than the test average might suggest. In addition, little could be said about students’ alternative beliefs.

On the other hand, a result of no correlation (or more realistically, a very low correlation), as illustrated in Figs. 1d and 2d, would prove that the two-item test does not measure students’ understanding of one aspect of Newton’s First Law. That is, regardless of whether the total average score is low (40%) or high (80%), getting one item either right or wrong is not associated with getting the other item right or wrong beyond the random chance level. There are two possible interpretations of a zero or low correlation result. If the distracters on the two test items were carefully chosen to capture students’ commonsense alternative beliefs, then one interpretation is that students do not have coherent concepts (either a “First-Law” concept or a coherent alternative belief like impetus)—their knowledge is in fragmented bits and pieces. (The possibility of stages of learning is discussed in the next section.) The other interpretation is that the students do have coherent concepts, but the test items do not measure this knowledge. In either case, it would be inappropriate to say that this hypothetical two-item test measures students’ understanding of a First-Law concept. It is also impossible to interpret the meaning of the average score on the test. When the correlation is zero (or very low), even a high average score (80%) does not necessarily indicate that the items measure an understanding of the concept. [The students in case (d) have exactly the same average test score of 80% for each item as in the case of (a), (b), and (c).] Therefore, even a high average score on the test does not indicate that the students are First-Law “Newtonian” thinkers. At best, a high-average score indicates that more students have learned the appropriate knowledge fragments. This is the type of results (low correlations) that we obtained for the items on the FCI.

This example illustrates that a correlation analysis of student responses to the FCI items can yield some information about student concepts because it does, in fact, take into account both the Newtonian and non-Newtonian response pattern of each student. Of course, this example is oversimplified. Many items are needed to measure any understanding, and we might expect a sample of students to have different degrees of understanding of the different conceptual dimensions of the force concept, which would lead to complex patterns of correlations between the items. It is fairly easy to interpret one correlation on a hypothetical, two-item test. But there are 29 items on the FCI test, so there are 406 possible correlations between pairs of items on the test. It would normally be very difficult to discern complex patterns of correlations by simple inspection (although in the case of the FCI, the fact that the correlations were low was obvious by inspection). Factor analysis is a mathematical technique for reducing a complex system of correlations to fewer dimensions (factors) by literally factoring the matrix of correlations. If the test items measure different conceptual dimensions of universal force concept (such as Newton’s First Law, Newton’s Second Law, and Newton’s Third Law), and the distracters include the majority of possible common-sense alternative beliefs, then we would expect all the test items to be correlated. In addition, the items that measure an understanding of Newton’s First Law would be more strongly correlated with each other than they would be with the items that measure Newton’s Second Law, and so on. If this were the case, then a (rotated) factor analysis would yield one factor of First Law items, another factor of the Second Law items, and so on. Moreover, each factor would account for a large proportion of the total variance in students’ scores. The guideline in test construction is that the significant factors should
account for at least one-half (50%) to two-thirds (67%) of the total variance in students' scores (Ref. 5, p. 180).

Our results indicated, however, that the FCI items are only weakly correlated. For the university sample, only 71 of the 406 correlations (17%) are above 0.19, and only 9 are above 0.30 (highest is 0.57). We had similar results for the high-school sample. Only 47 of the 406 correlations (12%) are above 0.19, and only 9 are above 0.30 (highest is 0.46). Since there are very few strong relationships (correlations) between the items, the factor analysis yielded one or two weak factors that accounted for only 16% (university sample) to 20% (high-school sample) of the total variance in students' scores. As we mentioned previously, there are two interpretations of this result. One interpretation is that students have coherent theories about forces (including alternative beliefs), but the FCI does not measure this knowledge. It seems more likely that the FCI measures bits and pieces of students' knowledge that do not necessarily form a coherent force concept.

As H&H point out, our results and interpretation are consistent with their previous analysis that students' force concepts appear to be vague and undifferentiated, loosely related, and sometimes inconsistent.25 We also agree with H&H that the six conceptual dimensions on Table I (Ref. 2) could be used as the Newtonian standard against which student concepts could be compared in detail. However, to find out if students actually interpret these concepts in the same way, an analysis similar to a factor analysis of students' scores is needed. The total FCI score does not necessarily measure "the disparity between student concepts and the Newtonian force concept." That is, a low score certainly shows a student does not have this concept, but our results indicate that high scores also do not show that students have a coherent force concept. This difference in our interpretation of the total FCI score is explained more fully in the following section.

What Does the FCI Score Tell Us?

In their original article, Hestenes and Wells10 presented a graph of the posttest FCI scores versus the Mechanics Baseline Test scores for 106 students in a calculus-based course for non-majors at Harvard University. The authors claim that the Baseline test measures the qualitative understanding of mechanics necessary for competence in problem solving (although they present no evidence that students' scores on the Baseline test correlate with their performance on physics problem-solving tests). Harvard University students who scored below 60% on FCI also scored below 60% on the Baseline test, so the authors suggest that a score of 60% on the Inventory is a "conceptual threshold" for problem-solving competence in physics. Below this threshold, the students' grasp of Newtonian concepts is "too limited for effective problem solving." Similarly, only students above 80% on the FCI were able to score above 80% on the Baseline test, so the authors suggest that the 80% score on the FCI is a "threshold" for mastery of basic Newtonian thinking. Although this is an interesting hypothesis, we did not mention these thresholds because the evidence the authors supplied did not support an interpretation of the word "threshold" to mean a qualitatively different way of thinking.

We could find these kinds of "thresholds" for any two variables that have a relatively high, but not perfect correlation. For example, we graphed the height versus weight of a sample of college men and found that "short" men (less than 66.5 inches tall) weighed less than 145 lbs, while only "tall" men (greater than 71 inches tall) weighed more than 210 lbs.11 This result does not, however, suggest that 66.5 inches and 71 inches are thresholds for qualitatively different growth mechanisms for short and tall men. In their response to our article, however, H&H suggest just this type of thresholds. They hypothesize that students learn Newtonian mechanics in stages. That is, learning is not just a matter of a monotonic increase in understanding of Newtonian mechanics, but qualitatively different kinds of thinking occur in different stages of learning. The students in Stage I (those who score less than a 60% on the FCI) do not have a coherent Newtonian force concept. Their thinking is characterized by undifferentiated, fragmented concepts about force and motion. Students in Stage II (those who score between 60% and 85% on the FCI) have developed a fairly coherent universal force concept. H&H are confident that students who score above an 85% on the FCI (Stage III) are "confirmed Newtonian thinkers." Furthermore, they claim that our results support this interpretation.

This support is apparently the fact that we did not find strong correlations and factor structure. H&H claim that the fact that our data do not cluster on conceptual dimensions "simply means that any 'Newtonian signal' they might contain is swamped by the noise of false positives." The false positives arise from the students who scored below 60%. Since the hypothesis is that these students do not have a coherent Newtonian force concept, the correlation between the items for these students should be very low. To extract the signal from the noise, the authors suggest that we perform a factor analysis on the group of students with the total FCI scores between 60% and 85%, and separately for the group with scores above 85%. Presumably, since the students who score between 60% and 85% have developed a fairly coherent Newtonian force concept, we should find fairly high correlations between the items and therefore a stronger factor structure. Since the students who score above 85% are hypothesized to be confirmed Newtonian thinkers, we should find very strong correlations between the items, and therefore a factor structure that matches the six conceptual dimensions on Table I of Ref. 2.

We had done this analysis before we published our article, but since we did not interpret "threshold" in the same way as the authors, we did not report the results of this analysis. We will correct this omission now. In our university sample of 512 students (FCI mean = 68%), 141 students scored below 60% (mean = 47%), 268 scored between 60% and 85% (mean = 71%), and 103 students scored above 85% on the FCI (mean = 90%). We found that the correlations between the items for the three samples are very similar to those of the whole sample, only somewhat lower. (Note: lower correla-
tions are expected when two items are linearly related, but the range or "lever arm" is artificially reduced by cutting off the high or low end of the sample.) For the below-60% sample, only 24 of the 406 correlations (6%) are above 0.19, and only 4 are above 0.30 (highest is 0.46). For the 60% to 85% sample, only 14 (3%) of the correlations are above 0.19, and only 5 are above 0.30 (highest is 0.55). The sample size of the above-85% sample is probably too small to compensate for the small variation in scores, so the statistical analysis should be interpreted with extreme caution. In fact, on four of the items the students scored 100%, but these items do not supply any information about whether the students have any unified force concepts. We found that the correlation matrix for the remaining 25 items is similar to the other two samples—only 22 of the 300 correlations (5%) are above 0.19, and only 6 of these are above 0.30 (highest is 0.40). Since there are very few strong relationships (correlations) between the items for each sample, the three factor analyses did not yield any factors that accounted for more than 8% to 9% of the total variance in the scores of the students in each sample.

Of the 179 students in the high-school sample (mean FCI = 51%), 117 students scored below 60% of the FCI (mean = 42%), 59 students scored between 60% and 85% on the FCI (mean = 67%), and only 3 students scored above 85%. We examined the correlations between the items for both the below-60% sample and 60% to 85% sample (although the size of the latter sample is very small for a statistical analysis, so the correlations should be interpreted with caution). We found that the correlations between items for the two samples are again essentially identical to those of the whole sample, only somewhat lower. For the below-60% sample, only 32 of the 406 correlations (8%) are above 0.19, and only 3 of these are above 0.30 (highest is 0.50). For the 60% to 85% sample, 66 of the correlations (16%) are above 0.19 (although only 9% are significantly different from a zero correlation), and 17 are above 0.30 (highest is 0.68). Since there are very few strong relationships (correlations) between the items for each sample, the two factor analyses did not yield any factors that accounted for more than 13% of the total variance in the scores of the students in each sample.

These results indicate that students who score between 60% and 85% on the FCI do not have a more coherent Newtonian force concept than the students who score below 60%. Therefore, our results do not support an "entry threshold to Newtonian physics" or the three-stage hypothesis of conceptual evolution in learning Newtonian mechanics. Our results also do not, as H&H claim, support their statement that the FCI measures "coherence across all dimensions of the Newtonian force concept." Moreover, we found that the type of instruction that university students received does not influence their response patterns to the FCI. Our university sample consisted of students from three different sections of the calculus-based course. In one section, the lectures and recitations were taught in the traditional manner. The average FCI posttest score was 63% and the pretest-to-posttest gain was 29% of the maximum possible gain (100 – pretest score). In the second section, the lectures were taught in a traditional manner, but during recitations students solved problems in cooperative groups. The average FCI posttest score for this section was 69% and the pretest-to-posttest gain was 37% of the maximum possible gain. In the third section, an explicit problem-solving strategy was taught, as well as the cooperative group problem-solving recitations. The average FCI posttest score was 72% and the pretest-to-posttest gain for this section was 49% of the maximum possible gain. Yet for the three sections, the correlations among the FCI items were very similar: only 4% to 7% of the 406 correlations were above 0.30. Since the correlations were weak, each of the factor analyses yielded only one weak factor which accounted for only 16% to 17% of the total variance in students' scores in each section. Since the results were the same for the three sections, we combined the sections for the factor analysis reported in our article.1

Using the Force Concept Inventory

We agree with H&H that if a test is a good predictor, then any uncertainty in the interpretation of the test scores is irrelevant to the prediction. In a previous publication,2 H&H found that the Mechanics Diagnostic Test (pretest) had a correlation of 0.56 with grades in introductory physics courses at Arizona State University. By combining the Mechanics Diagnostic Test with a simple math test, they were able to accurately predict course grades. We found, however, that the correlation between the FCI (pretest) and course grades was only 0.27 in each of three sections of calculus-based physics courses taught by different instructors with different instructional strategies at the University of Minnesota. Even combined with a math test, this correlation is too low to accurately identify students who are likely to have undue difficulty with the course. Consequently, we advise caution in using the FCI to make decisions about individuals before the accuracy of the prediction has been checked with the student population at the institution.

We also agree with H&H that there is a need for a "nationally normed" test to evaluate the effectiveness of instruction in introductory physics courses. The FCI, which has reasonable face and content validity, is the best test currently available, but we believe a more valid test from the point of view of students' responses is needed. For our large samples of students at the University of Minnesota and a suburban high school in Minnesota, our results indicate that regardless of whether students have higher (60% to 85%) or lower (below 60%) total scores on the FCI, and regardless of the type of instruction they received, the FCI items are only weakly related (correlated) to each other, and do not form strong factors. As we mentioned previously, there are two interpretations of this result. One interpretation is that students have coherent theories about forces (including alternative beliefs), but the FCI does not measure this knowledge. It seems more likely, however, that the FCI measures bits and pieces of students' knowledge that do not necessarily form a coherent force concept. In either case, instructors should be cautious about concluding that the FCI measures students' understanding of the Newtonian force concept. Even a high FCI posttest score (and a high gain) does not necessarily indicate
that students are developing a universal force concept, which is an essential goal of an introductory physics course. Based on our results, we do not believe urging caution in interpreting and using FCI scores is “ill-considered” advice. We reiterate that more research is needed to determine what the FCI is actually measuring.

References
11. Height and weight data are from laboratory sections of physics taught by Mark Hollabaugh at Normandale Community College in Bloomington, Minnesota.

Appendix: What Does the Factor Analysis Actually Measure?

We have explained that factor analysis is commonly used in educational measurement in the process of constructing and validating tests.4–7 It is only one of several types of evidence that can be used to help validate a test. We are pleased that Hestenes and Halloun pointed out Stephen Gould’s book, The Mismeasure of Man.12 Gould gives an excellent, non-technical explanation of factor analysis. We strongly recommend that readers interested in grasping the meaning of factor analysis read Chapter 6 of his book. Gould has no argument with the mathematical technique of factor analysis. He does point out the danger of misinterpreting strong correlations obtained by that analysis. Gould states on page 268 that:

Under certain circumstances, factors may be regarded as hypothetical causal influences underlying and determining the observed relationships [correlations] between a set of variables... My complaint lies with the practice of assuming that the mere existence of a factor, in itself, provides a license for causal speculation.

We concur with both Gould and H&H that there is “no guarantee whatsoever that factors can be given a sensible interpretation.” If we had found strong factors in our analysis and made causal claims, then we would be criticized for misusing statistics, as H&H imply in their Appendix. In fact, however, we found that the items on the FCI are only weakly correlated, so the factor analysis yielded one or two weak factors that accounted for only 16% (university sample) to 20% (high-school sample) of the total variance in students’ scores. We correctly concluded from this analysis that, from the viewpoint of the students taking the test, the items on the FCI are only loosely related to each other, and do not appear to measure a universal force concept. In other words, finding a strong factor structure does not necessarily imply causal relationships, but not finding a strong factor does imply that no such relationships exist.

H&H further criticize factor analysis because the test items are assumed to be linearly related to a set of uncorrelated factors, and a linear relationship “can be justified only in very special cases.” In fact, linear models are appropriate to use when data is expected to have monotonic behavior, especially when the data has a large scatter. For example, H&H use a linear model in their regression analysis to test whether the Mechanics Diagnostic Test, a math test, and previous courses in math and physics predict grades in a college physics course. In practice, if any linear model yields significant results, the model can then be refined by testing for non-linear contributions. There is a factor analytic technique that can be used to test for non-linear contributions. There is a factor analytic technique (called oblique rotations) to correct for the fact that factors may, in fact, be correlated. If we had found any indication of a factor structure to the data from our first linear model, we would have employed these techniques and reported the results. However, we found only weak correlations between the items and consequently one or two weak factors, so using these techniques was not needed (although we did, for the sake of completeness, actually apply these techniques and found no difference in the results).

Finally, H&H criticize the oversimplified example in our Appendix because we treat velocity and acceleration as independent variables, and they “are certainly not conceptually independent.” In fact, the use of velocity and acceleration in this hypothetical example was prompted by our analysis of the FCI items. The correlation between the velocity item (#20 on the FCI) and the acceleration item (#21) is only 0.24 for our university sample and 0.19 for our high-school sample. These are close to the correlations we used in our example. We did assume that it would be possible to construct two questions similar enough to items #20 and #21, so that the correlation between the two velocity items and the correlation between the two acceleration items would be much higher than the low correlations between the velocity and acceleration items. In any case, our hypothetical example was intended only to illustrate how to conduct a factor analysis, not how to construct and validate a set of real questions that test for velocity and acceleration.

"Interpreting the FCI: A Reply" VOL. 33, NOV. 1995 THE PHYSICS TEACHER 511