The search for conceptual coherence in FCI data

Ibrahim Halloun and David Hestenes
Department of Physics and Astronomy, Arizona State University

Douglas Huffman and Patricia Heller (H&H)[1] have raised concerns about the interpretation of the Force Concept Inventory(FCI)[2] based on a factor analysis of their own FCI data. They have further clarified their position[3] in response to our critique[4]. It is now possible to pinpoint more precisely just where we disagree. But first we would like to emphasize our broad agreement with their views on the FCI, as we understand them.

H&H accept our claims about the validity of the FCI as a test for the Newtonian force concept with our analysis of this complex concept into six conceptual dimensions. They agree that the FCI is the "best test currently available" to "evaluate the effectiveness of ... introductory physics courses"[3]. They have used it extensively in their own courses, and we note that the results indicate that their instruction is significantly more effective than instruction at other universities for which we have examined data.

We are also in general agreement with their call for *caution* in interpreting FCI results, especially for individual students. We have emphasized that the FCI questions are merely probes of student thinking, so, for individual students, FCI results must be supplemented by information from other sources to get a reliable profile of student understanding. We part company with H&H in some of their reasons for caution. Our response[4] to their initial article[1] was prompted by concern over feedback from many teachers who construed it as questioning the validity of the FCI as an evaluation instrument. We hope it is now clear that the "caution" urged by H&H refers to a much narrower issue. They claim to have analyzed FCI data from "the students' point of view." Though they do not explain precisely what they mean by this, we infer from examining their account that their analysis of FCI data aims to answer the question: *How coherent is the students' understanding of the Newtonian force concept?* This is a very good question, and we would like to be able to answer it for any given population of students. In this note, however, we only wish to explain why we think that the method employed by H&H cannot extract a correct answer from FCI data.

The FCI was expressly designed and validated to measure "the disparity between student concepts and the Newtonian Force Concept"[4]. The simplest index of this disparity is the *total FCI score* for an individual student or the mean score for a population. H&H question the reliability of this index, asserting that "a low [FCI] score certainly shows a student does not have

[the Newtonian force] concept, but our results indicate that *high scores* also do not show that students have a *coherent* force concept"[3] (our italics). H&H confess to being surprised by their results — with good reason, we think, because there is strong evidence to the contrary.

We agree that their data supports the conclusion that most students do not develop a coherent understanding of the Newtonian force concept in introductory physics. However, we contend that high FCI scores (>85%) are extremely unlikely without considerable coherence. We base this contention on extensive experience with such students, including exploratory analysis of their FCI responses, interviews, and comparison with other measures of their understanding, especially the Mechanics Baseline test[5]. To be sure, student FCI scores can be increased by telling them to memorize the correct answers. However, our research shows that such increases decay rapidly, probably in large part because the non-Newtonian alternatives on the FCI are so compelling to students that they tend to override rote responses. Only a coherent Newtonian understanding appears able to overcome this tendency.

We believe that the H&H claim to the contrary is based on a flawed statistical analysis. Recall that they reduce each question to a dichotomous variable, scored 1 for a right answer and 0 for a wrong one. Then they compare questions in pairs using the "*correlation coefficient*"

$$r = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}, \qquad (1)$$

with the meanings of *a, b, c, d,* given in Fig. 1. H&H interpret the *correlation r* as a measure of *conceptual coherence* in the response pairs. We reject this approach for several reasons:

(1) Established practice. While H&H are correct in asserting that factor analysis has been widely used in constructing and validating tests, they fail to note that this practice has been severely criticized in the literature, especially for the case of dichotomous variables[6]. However, we need not appeal to the opinions of experts to decide the matter, because we can evaluate the applicability of *r* ourselves.

(2) Statistical validity. Aside from *r*, there are many other statistical measures, or *statistics*, as they are often called, which can be used to analyze test data. The choice of an appropriate statistic depends on the question to be asked of the data. As H&H assert, their choice of the statistic is based on the following premise: "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 answer both items correctly, *while students who do not understand this aspect of Newton' s First Law would tend to consistently select the powerful but incorrect non-Newtonian distracters*" (Italics added). We maintain that the part in italics is inappropriate for a statistic intended to test for coherent understanding of Newtonian concepts, because correlation among non-Newtonian responses is irrelevant to such a test. Newtonian theory implies nothing about responses based on any non-Newtonian theory, be they coherent or not. We know full well that non-Newtonian thinkers occasionally give Newtonian responses, whereas Newtonian thinkers rarely give non-Newtonian responses. This important fact is not reflected in the choice of the *r* statistic. We believe that better statistics are available for analyzing FCI data, but this is not the place to explain our preference.

The fact that *r* does not discriminate between right and wrong answers is expressed by the invariance of its definition (1) under interchange of *a* and *d* and of *b* and *c*. The unacceptable implications of this property are most obvious when *r* is applied to data from a population of near-Newtonian thinkers. Following H&H[3], in Fig. 1 we display mock data on two questions about Newton' s First Law for a hypothetical population of 500. For perfect Newtonians we should get the results in Fig. 1a, but *r* has an indeterminate value, so it tells us nothing about the situation. To get the highest possible correlation $r = 1$, at least one person must get *both* questions wrong, as in Fig. 1b. Among Newtonians, however, it would be more likely for someone to "slip up" on only one of the questions. If everyone gets at least one of the questions right, *r* again has an indeterminate value, as in Fig. 1c. If a few people miss each of the questions but no one misses both, then the correlation will be small and negative, as in Fig. 1d. Comparing Figs. 1b and 1d, we see that a shift in the number of mistakes by just one person is sufficient to shift *r* from a very low value to a very high value. All together these examples show that *r* is not an appropriate statistical measure of conceptual coherence.

(3) Real data. As evidence for their claim that high FCI scores are *not* indicative of a coherent Newtonian force concept, H&H cite the low correlations in their data for 103 university students with scores over 85%. In particular, they tell us that "only 22 of the 300 correlations [which have definite values] are above 0.19"[3]. The suggestion is that these are the only pairs exhibiting significant (though weak) conceptual coherence. On the contrary, we have seen that the only way to get a positive correlation coefficient is for some students to get both questions wrong, so these 22 cases concern questions with which (perhaps only a few) students are having some difficulty. H&H[3] dismiss the 106 pairings with the four questions on which all the students have perfect scores as uninformative because, we suppose, the values of *r* are indeterminate (as in Figs. 1a

and 1c), though we think that many if not most of these cases indicate strong conceptual coherence, and it is the statistic *r* which is uninformative.

As an aid to interpreting the significance of FCI scores, we have proposed a *three-stage model of conceptual* ***evolution*** (misprinted as "evaluation" in Ref. 4) in learning Newtonian mechanics. This model is based, in part, on our understanding of student "misconceptions," and in part, on the logical structure of Newtonian theory which implies, for example, that an understanding of kinematics is a logical prerequisite to understanding Newton' s First and Second Laws. H&H claim that their data shows that there is no such natural progression in student learning. Indeed, their data does not show a tendency for *r*-correlations to increase with increasing FCI scores. But this is not evidence against our suggestion that there is a corresponding increase in conceptual coherence, because *r* does not measure that. Rather, we think, it is a reflection of the right-wrong symmetry of *r*, with the wrong answers dominating in the lowest range and the right answers dominating in the highest range. We agree with H&H that our model should be subjected to statistical analysis, but a much more sophisticated analysis will be necessary to reach reliable conclusions. We anticipate that the degree of coherence in posttest FCI scores will depend, in part, on the nature of instruction. We agree with H&H that, under some instruction, the increase in FCI scores is mainly due to an accumulation of "bits and pieces of knowledge" rather than a more coherent understanding. However, we also anticipate greater coherence among FCI answers under instruction that consistently achieves exceptionally large gains in FCI scores.

We share with H&H the belief that the development of a *coherent* force concept in students is "an essential goal of introductory physics instruction." To promote that development we need ways to test for such coherence. H&H suggest that we need a better test than the FCI. In the mean time, we think that better statistical analysis of FCI results will help.

		item # 6 wrong	item # 6 right
item #26	wrong	[a] 0	[b] 0
	right	[c] 0	[d] 500

(a) $r = \frac{0}{0}$

		item # 6 wrong	item # 6 right
item #26	wrong	[a] 1	[b] 0
	right	[c] 0	[d] 499

(b) $r = 1$

		item # 6 wrong	item # 6 right
item #26	wrong	[a] 0	[b] 1
	right	[c] 0	[d] 499

(c) $r = \frac{0}{0}$

		item # 6 wrong	item # 6 right
item #26	wrong	[a] 0	[b] 1
	right	[c] 1	[d] 498

(d) $r = -.002$

Figure 1: Different correlation values when an overwhelming majority of students answers two dichotomous items correctly

References

1. D. Huffman and P. Heller. “What does the Force Concept Inventory actually measure?”, *Phys.Teach.* **33** (3), 138–143 (1995).
2. D. Hestenes, M. Wells, and G. Swackhamer. “Force Concept Inventory”, *Phys.Teach.* **30** (4), 141-151 (1992).
3. P. Heller and D. Huffman . “Interpreting the Force Concept Inventory. A reply to Hestenes and Halloun”, *Phys.Teach.* **33** (8), 504, 507–511 (1995).
4. D. Hestenes & I. Halloun. “Interpreting the Force Concept Inventory”, *Phys.Teach.* **33** (8), 502, 504–506 (1995).
5. D. Hestenes, and M. Wells. “A Mechanics Baseline Test”, *Phys.Teach.* **30** (4), 159-166 (1992).
6. See, for example: R. J. Mislevy. “Recent developments in the factor analysis of categorical variables”, *J. Ed. Statistics.* **11** (1), 3–31 (1986).