The puzzling reliability of the Force Concept Inventory

Nathaniel Lasry
Center for the Study of Learning and Performance, Concordia University, Montreal, Canada H3G 2V8 and John Abbott College, Department of Physics, Montreal, Canada H9X 3L9

Steven Rosenfield and Helena Dedic
Center for the Study of Learning and Performance, Concordia University, Montreal, Canada H3G 2V8 and Vanier College, Department of Physics, Montreal, Canada H4L 3X9

Ariel Dahan
Center for the Study of Learning and Performance, Concordia University, Montreal, Canada H3G 2V8 and Faculty of Science, McGill University, Montreal, Canada H3A 2T6

Orad Reshef
School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts 02138

(Received 8 November 2010; accepted 27 May 2011)

The Force Concept Inventory (FCI) has influenced the development of many research-based pedagogies. However, no data exists on the FCI's internal consistency or test-retest reliability. The FCI was administered twice to one hundred students during the first week of classes in an electricity and magnetism course with no review of mechanics between test administrations. High Kuder–Richardson reliability coefficient values, which estimate the average correlation of scores obtained on all possible halves of the test, suggest strong internal consistency. However, 31% of the responses changed from test to retest, suggesting weak reliability for individual questions. A chi-square analysis shows that change in responses was neither consistent nor completely random. The puzzling conclusion is that although individual FCI responses are not reliable, the FCI total score is highly reliable.

[DOI: 10.1119/1.3602073]

I. INTRODUCTION

The Force Concept Inventory (FCI) is a multiple-choice instrument[1] that asks conceptual physics questions using everyday terms. Answering does not involve calculations.[1–3] What instructors think that their students understand, and what the FCI results show, can be very different.[4] This difference has helped to make the FCI one of the most widely used instruments in physics education.[5]

In a study of more than 6500 college and university physics students, Hake[6] used changes in FCI scores before and after a semester of instruction to show that traditional methods of instruction were ineffective in altering students' preconceptions. Since Hake's study, the FCI has influenced the development of innovative pedagogies and has played a key role in facilitating acceptance by mainstream physics faculty members of many research-based pedagogies.

Although the FCI has been given more than one hundred thousand times[7] at several hundred institutions worldwide, little data exists on its reliability. In this paper, we study the reliability of the FCI and report some puzzling findings.

II. THEORETICAL FRAMEWORK

There are three main ways to measure the reliability of a test: Equivalent form reliability, internal consistency reliability, and test–retest reliability.

Equivalent form reliability is measured by the correlation of the score on an instrument with the score of the same group of students on a second instrument that measures the same construct. Two instruments that are frequently used to assess students' understanding of the concept of force are the FCI and the Force and Motion Conceptual Evaluation (FMCE).[9] The correlation between the FCI and FMCE has previously been determined in 2000 cases. A reasonably high correlation of $r = 0.78$ was found, indicating that the score on one instrument is a good predictor of the score on the other.[10]

Internal consistency reliability is measured by correlating the total scores on two distinct halves of a test to establish whether the sub-parts are consistent with each other. Internal consistency reliability assumes that a test measures a unique construct across all test items.[8] For instance, a question assessing prejudice against foreign cars could be "Are foreign red cars ugly?" This question assesses appreciation of foreign cars but is confounded with color (error). If the errors in answers to individual questions are random, we expect that the score obtained on any half of the test should be roughly the same as the score obtained on the other half. However, the correlation between the scores on two distinct halves of the test might depend on how the test is split. For instance, the correlation of scores on the first and second halves of the test may be different from the correlation between scores on odd and even numbered questions. To sidestep this issue, Kuder and Richardson developed a reliability coefficient, KR-20, which estimates the average correlation between all possible halves, provided all items in the test are scored as dichotomous variables (for example 0 for incorrect and 1 for correct).[8] The value of KR-20 varies between 0 (no internal consistency) and 1 (perfect internal consistency). Comparing scores between groups requires a minimum reliability coefficient of 0.70. Below this value, the instrument is usually considered to be internally inconsistent for research purposes.[11] Comparing scores between individuals requires KR-20 > 0.80.[11]

The internal reliability of the Mechanics Diagnostic Test was previously determined, and KR$-$20 was found to equal 0.86 for the pretest and 0.89 for the post-test.[2] To our knowledge, no assessment of the internal consistency of the FCI has been reported.

Test-retest reliability assesses the stability of test scores by administering a test twice to the same group of students. When measuring test–retest reliability, the lapse of time between test administrations should be long enough for students not to recall their answers, yet short enough for their knowledge about the topic to remain unchanged. There are two ways to measure test–retest reliability. For the total score to be considered reliable, the difference in the total score on the two test administrations should be statistically insignificant.[8] Alternatively, test–retest reliability is measured by correlating the scores students obtained on the test with their scores on the retest. As with internal consistency, a minimum correlation of 0.70 is required for research purposes. A correlation of above 0.80 is usually sought to infer the stability of scores.[11] To our knowledge, no test re-test reliability measurement has been done on any physics education instrument.

III. METHOD

The students in this study were enrolled in an electricity and magnetism course at John Abbott College, a publicly funded two-year college in Montreal, Canada. Students in the three cohorts studied (Fall 2007, $N = 37$; Fall 2008, $N = 38$; and Fall 2009, $N = 36$) had the same instructor (NL), textbook, laboratory components, and other course materials.

All students ($N = 111$) were given the FCI in the first week of the course. Students did not know in advance that they would be retested. Within a week from the first administration, students were offered bonus points for retaking the FCI. A total of 100 students (Fall 2007, $N = 31$, Fall 2008, $N = 34$, and Fall 2009, $N = 35$) were tested and then retested on the FCI. No mechanics instruction or review was given to these students between the two FCI administrations.

IV. RESULTS

We assessed internal consistency by computing KR$-$20 for the FCI test, retest, and combined data from both. Our results are shown in Table I. The results indicate that the FCI is internally consistent. That is, the scores on different halves of the FCI are highly correlated.

The average total score and the correlation between the test and the retest are shown in Table II. The correlation between test and retest total scores is $r = 0.89$. The difference in mean scores between the two test administrations, μ, is 1.6% with a sample standard deviation of 8.5%.

Assuming that changes in the total score are normally distributed, a z-test[12] was used to test the null hypothesis, $\mu = 0$, against the alternative hypothesis, $\mu \geq 0$. It was determined that μ is sufficiently small that we should regard the change in the average total score as due to random chance.

To examine the stability of responses to individual FCI questions, we categorized the transitions in responses to individual FCI questions between the test and the retest as either right-to-right, right-to-wrong, wrong-to-right, wrong-to-same-wrong, and wrong-to-different-wrong. The proportion of students' transitions in each category is shown in Fig. 1.

Table I. Internal consistency as determined by the value of KR$-$20.

Test	0.900
Retest	0.812
Test–retest Combined	0.865

Table II. Test, retest scores, and correlation.

Test–retest correlation	$r = 0.89$
Average total test score	46.9%
Average total retest score	48.5%

We further examined the average number of changes in answers. Of the 30 FCI questions, the average number of changes per student was 9.39, with a sample standard deviation of 4.30. Assuming a normal distribution of the number of changes, the 90% confidence interval is [2.31, 16.47] and the 95% confidence interval is [0.96, 17.82], which means that for 90% (95%) of students the average number of changes is between 2.31 (0.96) and 16.47 (17.82). These results suggest that it is unlikely (only 5%–10% chance) that a student will not change at least one or two answers in a test–retest situation.

The finding that the total score is reliable, although individuals on average change close to a third of their answers, is puzzling. To investigate this seemingly anomalous result further we developed a probability model for transitions.

V. PROBABILITY MODEL FOR TRANSITIONS

The first model we tested was whether the FCI is completely reliable. This model assumes that any response given on the test will be identical to the response given on the retest. Hence, the probability of right–right is 1, as is the probability of wrong-to-same-wrong. All other probabilities are zero. By multiplying the total number of transitions by their

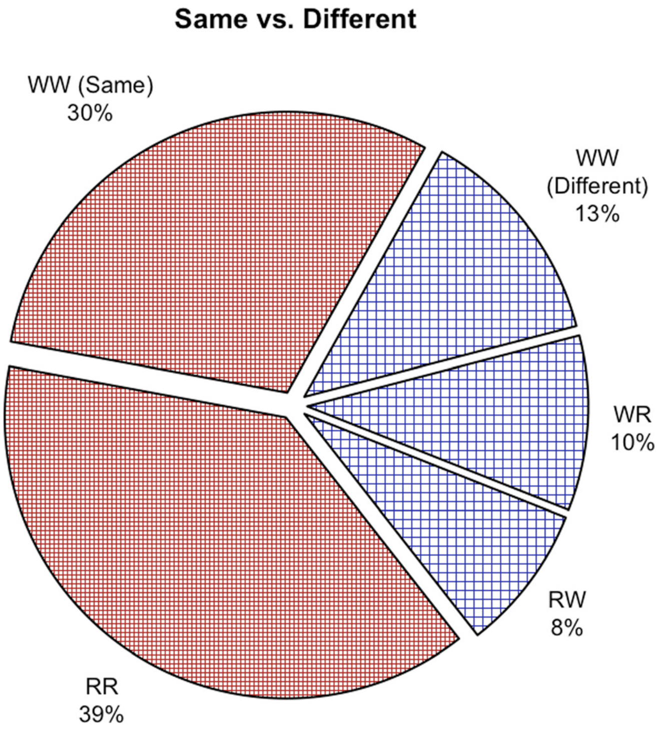

Fig. 1. FCI test–retest transitions: same versus different. Responses not changing between the test and the retest (same) are shown filled with a small crosshatch pattern. Responses that changed are filled with a large crosshatch pattern. The notation in the figure represents right-to-right (RR), right-to-wrong (RW), wrong-to-right (WR), and wrong-to-wrong (WW).

Table III. The transition probabilities for the hypothesis that the FCI is completely unreliable.

Symbol	Transition	Probability
P_1	right–right	0.2
P_2	right–wrong	0.8 $(1-P_1)$
P_3	wrong–right	0.2
P_4	wrong-to-same-wrong	0.2
P_5	wrong-to-different-wrong	0.6 $(1-P_3-P_4)$

probabilities, we obtained the expected frequencies for all transitions. We compared the expected with the observed frequencies to calculate $\chi^2 = 340.9$ ($p < 0.001$), which allows us to reject the hypothesis that the FCI is completely reliable.

The second probability model assumes that the FCI is completely unreliable. That is, students are equally likely to choose any answer on the retest, regardless of what was chosen on the initial test. For example, given that there is one correct answer and four incorrect answers, the probability for right–right is 0.2, and the probability for right–wrong is 0.8. Table III lists the probabilities of all possible transitions.

If we compare the expected and observed frequencies for this model, we obtain $\chi^2 = 4811.9$ ($p < 0.0001$), allowing us to reject the hypothesis that the FCI is completely unreliable.

The second model was iteratively corrected until the predicted transition frequencies optimally matched those observed. We tested values of P_1 between 0.1 and 0.9, with values of P_3 and P_4 with similar ranges. Because there are only three independent probabilities, the values of P_2 and P_5 were calculated from P_1, P_3, and P_4 (see Table III). For the minimum χ^2 value of $\chi^2 = 0.002$ we found that $P_1 = 0.82$ and hence $P_2 = 0.18$. Also, $P_4 = 0.57$, $P_3 = 0.19$, and $P_5 = 0.24$.

This probabilistic model shows the similarity between the probability of right–wrong ($P_2 = 0.18$) and of wrong–right ($P_3 = 0.19$). These probabilities can be used to account for the changes in total score as:

$$\text{change in total score} = P_3 \times (1 - \text{test score}) - P_2 \times \text{test score} \quad (1a)$$

$$= 0.19 \times 53.1\% - 0.18 \times 46.9\% = 1.65\%. \quad (1b)$$

The similarity of P_2 and P_3, combined with the fact that the average FCI score is very close to 50%, accounts for the small change in the total score. Note that the change in the total score is identical to the change in scores reported in Table II.

Like all measurements, the FCI total score is subject to error. This error consists of noise in student understanding and noise in the instrument. Given that only wrong–right and right–wrong change the total score, we can use them to estimate this error. We calculate the error in the total score over the test and retest as

$$\text{Error} = [P_3 \times (1 - \text{test score}) + P2 \times \text{test score}]/2 \quad (2a)$$

$$= (0.19 \times 53.1\% + 0.18 \times 46.9\%)/2$$
$$= 9.27\%. \quad (2b)$$

VI. DISCUSSION

Huffman and Heller[16] asked: "what does the FCI actually measure?" Using classical exploratory factor analysis, they examined whether there are groups of questions in the FCI that correlate with each other, which would indicate that those items measure the same idea. Their finding that the FCI questions correlated loosely led them to conclude that the FCI does not measure a single construct. Halloun and Hestenes[17] objected to the methodology used by Huffman and Heller and asserted that "the FCI score is a measure of one's understanding of the Newtonian concept of force." Halloun and Hestenes argued that if the students' understanding of force is not complete, then there is no reason for questions to correlate. Using a different methodology, our results show that the FCI has a high internal consistency reliability (KR-20 > 0.8) and hence support the notion that the total FCI score measures a unique construct. However, our analysis did not determine what this unique construct is. Given that Newtonian thinkers are likely to obtain a high FCI score, Halloun and Hestenes interpret the score as a measure of an understanding of the Newtonian concept-of-force.[17]

The high correlation between test and retest scores and the insignificant difference between the mean scores shows the stability of the total FCI score. Given the lack of instruction and the short time between the test and retest, the unique construct measured by the FCI should remain unchanged. Hence, the total FCI score is a reliable measure of the concept-of-force. Given the constancy of the total score, we might expect that the students would not change their answers often. However, we were puzzled to find that 31% of all responses were changed between the test and retest. Of the total number of responses changed, 13% did not affect the score, 8% decreased the score, and 10% increased the score.

From the perspective of a resources model,[13] the FCI questions provide a context that activates concept-of-force related schema or a related set of resources.[13,14] Given that the context for the test and retest was similar, the resources activated should be similar, and hence the probability of selecting a given FCI response should be similar. This similarity means that the probability of choosing an answer will be the same every time, not that they will choose the same answer every time. Hence, although individual responses fluctuate, the overall time-averaged mean-score is unchanged. In retrospect, our data provide good empirical support for the resource model.[13–15]

In their rebuttal to Huffman and Heller[16] and Halloun and Hestenes[17] assert that the FCI has measurement error, as do all tests. A false-positive occurs when a non-Newtonian thinker selects a correct response. For instance, in think-aloud student-interviews, Thornton *et al*. showed that some students chose a correct FCI response using incorrect reasoning.[10] In a previous study using latent Markov chain modeling,[18] we found statistical evidence of the same false-positive (FCI question 16) reported in the qualitative analysis in Ref. 10. Conversely, a false-negative occurs when a Newtonian thinker chooses an incorrect response.

A right–wrong transition could be an indication of either a false-positive on the test or a false-negative on the retest. Similarly, a wrong–right transition could be an indication of either a false-negative on the test or a false-positive on the retest. We can interpret Eq. (2) as the average total of false-positives and false-negatives on either test. This interpretation of Eq. (2) assumes that the occurrence of false-positives or false-negatives does not change between test and retest. How does this error differ from the average change in score between the test and the retest reported in Eq. (1)?

The difference between both errors reduces to the difference between precision and accuracy. The FCI is precise because its total score has an error of 1.6%. Its accuracy is 9.3%. The analogy can be made with a meter-long stick being used to measure a yard. Repeated measurements using this stick to measure yards would be very reproducible (precise) but inaccurate by roughly 10%.

Our students have an average FCI score below 50%, indicating that these students are not likely to be Newtonian thinkers. Pre-Newtonian thinkers are thought to have conceptions that are not fully coherent or consistent.[19] Our data show that the probability of maintaining a correct answer is 82%, and the probability of maintaining the same incorrect answer is 57%. Inconsistency appears to lie mostly in incorrect responses.

We also found a 24% chance of changing an incorrect answer to a different incorrect answer. These changes do not affect the total score and hence are not part of our error calculations. Given that roughly one of four wrong answers is changed, these changes warrant further investigation.

VII. CONCLUSION

This study confirms that the FCI total score reliably measures a single concept, although our analysis is silent as to the nature of this concept. High test–retest reliability shows that FCI total score is a precise metric.

A different picture emerges when examining individual questions. We found that although the macro conceptual state gauged by the FCI total score is unchanged, responses to individual questions are not, and roughly one third of responses are changed between test and retest (see Fig. 1). This finding can be seen as providing empirical support in favor of the resources framework.[13–15]

Our results are based on a small sample of students enrolled in a Canadian two-year college. Although the sample size of $N = 100$ was sufficient for the statistical methodologies employed, it would be useful to see this experiment replicated with more subjects and across multiple populations.

[1] D. Hestenes, M. Wells, and G. Swackhamer, "Force concept inventory," Phys. Teach. **30**, 141–158 (1992).

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

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

[4] E Mazur, *Peer Instruction: A User's Manual* (Prentice Hall, Upper Saddle River, NJ, 1997).

[5] L. C. McDermott and E. F. Redish, "Resource letter: PER-1: Physics education research," Am. J. Phys. **67**, 755–767 (1999).

[6] R. R. Hake, "Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses," Am. J. Phys. **66**, 64–74 (1998).

[7] The order of magnitude of the FCI data at Arizona State University is at least 50000 cases. Through the Interactive Learning Toolkit test module, the Mazur group at Harvard University has close to 10000 cases. There are several hundreds of undocumented FCI administrations each semester worldwide. We would not be surprised if the order of magnitude of FCI administrations was close to one million since its initial publication in 1992.

[8] R. P. McDonald, *Test Theory: A Unified Treatment* (Lawrence Erlbaum Associates, Mahwah, NJ, 1999).

[9] R. K. Thornton and D. R. Sokoloff, "Assessing student learning of Newton's laws: The force and motion conceptual evaluation and the evaluation of active learning laboratory and lecture curricula," Am. J. Phys. **66**, 338–351 (1998).

[10] R. K. Thornton, D. Kuhl, K. Cummings, and J. Marx, "Comparing the force and motion conceptual evaluation and the force concept inventory," Phys. Rev. STPhys. Educ. Res. **5**, 010105 (2009).

[11] J. M. Bland and D. G. Altman, "Cronbach's alpha," Br. Med. J. **314**, 572 (1997).

[12] R. C. Sprinthall, *Basic Statistical Analysis* 7th ed. (Allyn and Bacon, Boston, 2003). Also see <en.wikipedia.org/wiki/Z-test>.

[13] D. Hammer, A. Elby, R. E. Scherr, and E. F. Redish, "Resources, framing, and transfer," in *Transfer of Learning from a Modern Multidisciplinary Perspective*, edited by J. Mestre (Information Age Publishing, Greenwich, CT, 2005)

[14] E. F. Redish, "A theoretical framework for physics education research: Modeling student thinking," in *Proceedings of the 2004 Enrico Fermi Summer School*, Course CLVI, Italian Physical Society, edited by E. Redish, C. Tarsitani, and M. Vicentini, pp. 1–63.

[15] M. S. Sabella and E. F. Redish, "Knowledge organization and activation in physics problem solving," Am. J. Phys. **75**, 1017–1029 (2007).

[16] D. Huffman and P. Heller, "What does the force concept inventory actually measure?," Phys. Teach. **33**, 138–143 (1995).

[17] D. Hestenes and I. Halloun, "Interpreting the Force Concept Inventory: A response to March 1995 Critique by Huffman and Heller," Phys. Teach. **33**, 502–504 (1995).

[18] H. Dedic, S. Rosenfield, and N. Lasry, "Are all wrong FCI answers equivalent?," AIP Conf. Proc. 1289, 125–128 (2010).

[19] S. Vosniadou, "Capturing and modeling the process of conceptual change," Learn. Instr. **4**, 45–69 (1994).