

Common Concerns About the Force Concept Inventory

Charles Henderson

The Force Concept Inventory (FCI) is currently the most widely used assessment instrument of student understanding of mechanics.¹ This 30-item multiple-choice test has been very valuable to the physics education community by helping to show that students can solve common types of quantitative problems without a basic understanding of the concepts that are involved.² Since the test is so easy and quick to administer, many physics instructors have given it to their classes and have been surprised by the low scores of their students. This has, in part, helped to fuel the growing interest in physics education research. Although there have been some concerns about how to interpret FCI scores in terms of relating them to deep-seated student misconceptions or coherent ideas of force,³ the value of the test for helping to evaluate the effect of curricular changes is widely accepted. One of the most useful things about the FCI is that there are national data available that show test results do depend, to a large degree, on instructional practices and

give curricular reformers a scale upon which to measure their success.⁴

At the University of Minnesota we have been using the FCI since 1993 to gather data about our introductory calculus-based physics courses. The University of Minnesota is in a rather unique position in that we have been modifying the curriculum of all five sections of introductory calculus-based physics (rather than a single experimental section, as is commonly done at other universities). Because of the large scale and duration of our reform efforts, we are accountable to many faculty members and we frequently make use of FCI scores in describing course outcomes. We find that faculty members often have questions concerning the administration and interpretation of the FCI. We are in a unique position to address many of these questions since we have an unusually large amount of student FCI data that come from students who were all given fairly similar instruction. This paper will address the following questions:



Charles
Henderson

Charles Henderson is an assistant professor of physics at Western Michigan University. He has taught physics in a variety of settings including high school, community college, four-year college, and research university. Henderson's research interests lie in the area of physics education, particularly in understanding physics teachers' beliefs and values about teaching and learning. Such research will help improve physics teaching by facilitating communication between the physics education research community and the larger community of physics teachers. The data for this paper were collected while he was a doctoral student with the Physics Education Research Group at the University of Minnesota.

*Western Michigan University
Kalamazoo, MI 49008-5252;
Charles.Henderson@wmich.edu*

1. Can the FCI be used as a placement test?
2. How do we know that students take the FCI seriously when it is not graded?
3. Doesn't giving the FCI as a pre-test influence the post-test results?

All of the data gathered for this paper are from the introductory calculus-based physics courses at the University of Minnesota, and it is important for readers from other institutions to keep in mind that results obtained in their instructional settings might not be similar to those presented in this paper (as mentioned previously, it is well known that FCI scores depend on instructional practices).

Setting

About 850 students enroll in introductory calculus-based physics each fall semester at the University of Minnesota. The goal of our course is to have students learn physics by solving problems. We attempt to accomplish this goal through an instructional practice known as cooperative group problem solving that has been described elsewhere.⁵

Students select one of five lecture sections meeting at various times during the day. Each lecture section has between 80 and 250 students. The course follows a traditional model of three lecture hours, two lab hours, and one recitation hour each week. The focus of the lectures varies somewhat with the lecturer; however, all labs and recitation sessions use cooperative group problem solving. Each lecturer is responsible for writing quizzes during the semester, and all five lecturers collaborate to write a common final exam.

FCI Testing

The first thing that students do during the first lab session is to take the FCI as a pre-test. The test is offered to all students who come to the first lab. Students are told that the test is voluntary and that their participation in the testing will not affect their course grade. The post-test has been given on the final exam (1997 and 1998) or during the last lab session (1999). When it was given on the final exam, it counted toward the student's final exam grade. When it was given in the lab, it was voluntary.

At the end of the semester, students' pre-test and post-test scores are matched. Student scores are elimi-

nated from this study if they left 20% (six questions) or more of the FCI items blank on either the pre- or post-test.

There are two versions of the FCI. We began using the revised 30-item version of the FCI⁶ in 1997. To avoid possible problems involved with comparing scores on the original and revised FCI, only scores from the revised FCI are used in this paper.

Can the FCI Be Used as a Placement Test?

Some universities have placement tests that physics students must take in order to decide what level of physics course would be most appropriate for them. The goal of giving a placement test is to identify students who are very likely to do poorly in a given class and suggest alternative or supplementary classes for them to take. A placement test is effective if it is able to distinguish between students who will do well in the class and students who will do poorly in the class.

In order to determine whether this is an appropriate role for the FCI in our introductory calculus-based physics course, we looked at the relationship between FCI pre-test scores and success in the course. Based on their final grades, students were put into one of six grade categories (A, B, C, D, F-I-W, drop), where students who fail the class, take an incomplete, or withdraw are lumped together. In addition to the letter grades, there were also some students who took the FCI pre-test but subsequently dropped the course (dropping is different from withdrawing in that it occurs earlier in the term and there is no record placed on a student's transcript). These students who dropped the course could have done so because they were doing poorly or for other reasons. For the purposes of interpreting our data, we have assumed that a student who gets an A, B, or C in the class has been successful and that a student who gets a D, F, I, or W, or drops the class, has been unsuccessful.

Figure 1 breaks the class into 10 groups based on the FCI pre-test score. In each FCI pre-test group, the percentage of students who fall into each of the six grade categories is shown. As you can see from the graph, the FCI pre-test can do a reasonable job of predicting success in the class (almost all students, 94%, who scored 19 or higher on the pre-test were successful in the class, most getting an A or B), but does not

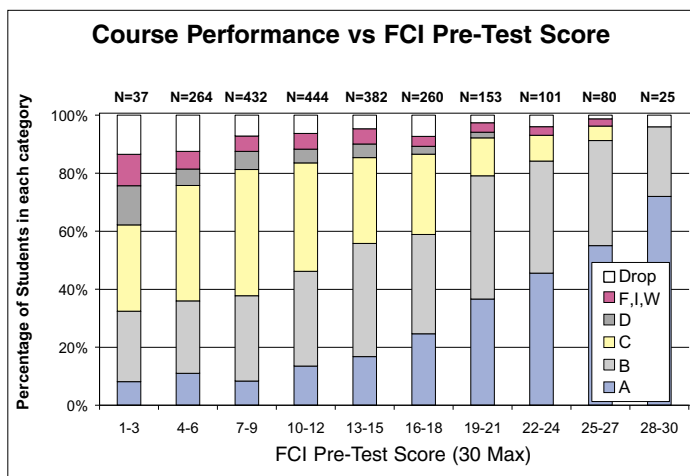


Fig. 1. Student course performance as a function of FCI pre-test score. All student pre-test data from fall 1997, fall 1998, and fall 1999 ($N = 2178$).

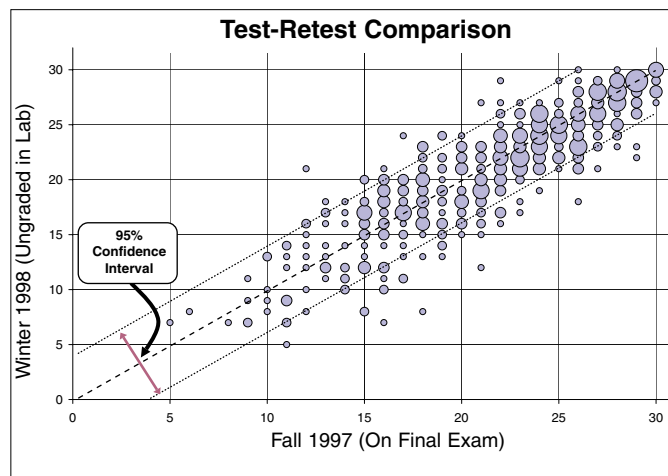


Fig. 2. The relationship between each student's ($N = 500$) score on the graded FCI on the fall 1997 final exam and his/her score on the ungraded FCI at the beginning of winter 1998 (three weeks later). The number of students in each group is represented by the dot size.

do a good job of predicting failure in the class (even in the lowest FCI pre-test score group, more than 60% of the students were successful in the class). Regardless of what cutoff score might be chosen, if we were to use the FCI as a placement test, many students would be inappropriately advised not to take our introductory calculus-based physics course.

How Do We Know that Students Take the FCI Seriously When It Is Not Graded?

Our FCI pre-tests are always ungraded and sometimes the FCI post-tests are ungraded. Since students are not penalized for not taking the test seriously, the question naturally arises as to how meaningful these test scores are. If we are to treat these ungraded scores as meaningful, we must determine to what extent an ungraded FCI test represents a student's best work. This question was examined using two different methods.

Examining student answer patterns for signs of lack of seriousness

There are several types of student answer patterns that may indicate that a student is not taking the FCI seriously. The five types of patterns we looked for were (1) refusing to take the test; (2) drawing a picture on the Scantron answer sheet; (3) answering all A's, B's, etc.; (4) leaving six or more blanks; and (5) other

patterns (such as ABCDE, EDCBA, AAAAA, BBBB, etc.) any place in their responses. Table I shows the percentage of students who fall into each of these categories based on the conditions under which the student took the FCI. We might expect to find differences between the pre-test and the post-test based on students' different knowledge, but the chart shows that for the ungraded pre-test and the ungraded post-test, the percentage of students in each of these categories is very similar. The main differences on the chart depend on the conditions under which the test was administered — i.e., whether or not it was graded. To determine the maximum percentage of students not taking the test seriously when the FCI is given ungraded, we found the difference between the percentages of students in each group when the FCI was given graded and when it was given ungraded.

Using this method, we can estimate that the maximum percentage of students who don't take the FCI seriously when it is ungraded is 2.8%. Thus, the response patterns of our students on the FCI indicate that almost all of them are taking the test seriously. Further, since it is relatively easy to identify students who refuse to take the test (0.5%) or leave a lot of blanks (1.4%), these 1.9% of students can be eliminated from the sample, leaving at most 0.9% of students who might have lower scores on ungraded tests due to lack of seriousness.

Table I. Signs of lack of seriousness in student answer patterns on the FCI.

	Pre-Ungraded	Post-Ungraded	Post-Graded	Maximum % of students not taking test seriously as a result of grading option
	<i>N</i> = 1856	<i>N</i> = 524	<i>N</i> = 1332	
	1997, 1998, 1999	1999	1997, 1998	
Refuses to take test	0.5%	0.5%	0.0%	0.5%
Draws a picture	0.2%	0.2%	0.0%	0.2%
Answers all A's, B's, etc.	0.0%	0.0%	0.0%	0.0%
Leaves a lot of blanks (six or more)	1.5%	1.5%	0.1%	1.4%
Other Patterns				
ABCDE, EDCBA	0.8%	1.0%	0.5%	0.5%
Six A's, B's, etc.	0.2%	0.2%	0.0%	0.2%
Total	3.2%	3.4%	0.6%	2.8%

Giving the same group of students the FCI twice, in a graded and ungraded situation, where it is unlikely that they learned new physics in between

The students identified in the previous section were ones who obviously did not take the test seriously. There may, however, be students who, when the test is ungraded, just don't bother to think about their responses as carefully as they would have had the test been graded. It is plausible that such a lapse in careful thought would decrease a student's test score. We wanted to find out if this effect exists and, if so, how large an effect it is.

In the fall of 1997, students took the FCI on the final exam of their first quarter of introductory physics. Three weeks later, during the first week of the second quarter of physics, in the winter of 1998, students were asked to take the FCI as an ungraded test during the first lab session. Since students were done with their physics course and on winter break during the intervening three weeks, it is unlikely that they attempted to learn any new physics in between the two administrations of the test. Figure 2 shows the relationship between the graded and ungraded tests. As you can see there is a high correlation ($r = 0.88$) between the two sets of scores. Some students had higher scores when the test was given ungraded, some students had lower scores, and some students received the same score. There is a line on the graph that

shows the expected result that a student's score on both tests would be the same. Surrounding this line are two parallel lines representing the 95% confidence interval of an individual student's FCI score (4.0 items) as determined by a separate analysis of FCI scores based on a measurement of the reliability of the test.⁷ Since most of the data points fall within this band, it is clear that most of the deviation in test scores can be attributed to measurement error rather than to the conditions of testing.

There does, however, appear to be a small effect due to the conditions of testing. Comparing the average of each group, we find a difference of about half of an FCI item between the graded test (21.4 ± 0.2) and the ungraded test (20.9 ± 0.2). This difference is statistically significant at the 5% level on a matched sample t-test. This half of an FCI item can then be considered to be the maximum difference in FCI scores that can be attributed to lack of seriousness. We think of this as a maximum since other factors, such as forgetting material over winter break, might also lead to a decrease in score on the ungraded test. We don't consider this potential half-item decrease in the class average to be of much concern since it is approximately the same as the statistical uncertainty in the FCI average for a class of 100 students.⁸

Combining these two methods of looking for lack of seriousness on ungraded tests, we can say that there is strong evidence that almost all of our students take

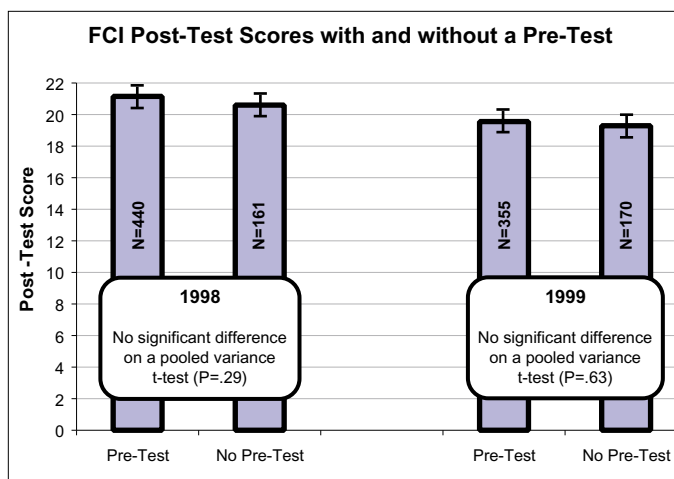


Fig. 3. A comparison of FCI post-test scores for groups of students who did and did not take the pre-test.

the test seriously when it is not graded and that there do not appear to be substantial problems in comparing results of graded and ungraded tests.

Doesn't Giving the FCI as a Pre-Test Influence the Post-Test Results?

A common concern about giving the FCI as both a pre-test and a post-test is that students' post-test scores might be inflated because students have already been exposed to the material on the pre-test. There are a number of possible reasons why this might occur. By taking the pre-test, for example, students may be sensitized to certain topics and then pay closer attention to these topics when they come up in the class. On the other hand, the pre-test is taken very early in the semester and students have no idea that they will ever see the same test again. We decided to see if taking the FCI as a pre-test has any influence on post-test scores. In two years (1998 and 1999), approximately one-quarter of the students were not given the FCI as a pre-test. In 1998 these students were given a different conceptual test (the TUGK),⁹ and in 1999 these students were not given any conceptual test. As you can see from Fig. 3, there are no statistically significant differences in post-test scores between these two groups. Thus, taking a pre-test does not appear to bias post-test results.

Conclusions

Because of the large number of students who take the FCI each year at the University of Minnesota, we have been able to address some common concerns about using the FCI. Data presented have shown that for students at the University of Minnesota:

1. The FCI is not appropriate for use as a placement test.¹⁰
2. There is little difference between FCI scores when the test is given graded versus ungraded.
3. Giving the FCI as a pre-test does not affect the post-test results.

References

1. Lillian C. McDermott and Edward F. Redish, "Resource Letter PER-1: Physics education research," *Am. J. Phys.* **67**, 755–767 (Sept. 1999).
2. See, for example, Eric Mazur, *Peer Instruction: A User's Manual* (Prentice Hall, Upper Saddle River, NJ, 1997).
3. Several papers have been published in *The Physics Teacher* regarding the interpretation of the results of the FCI: Richard Steinberg and Mel Sabella, "Performance on multiple-choice diagnostics and complementary exam problems," *Phys. Teach.* **35**, 150–155 (March 1997); David Hestenes and Ibrahim Halloun, "Interpreting the Force Concept Inventory: A response to March 1995 critique by Huffman and Heller," *Phys. Teach.* **33**, 502, 504–506 (Nov. 1995); Pat Heller and Doug Huffman, "Interpreting the Force Concept Inventory: A reply to Hestenes and Halloun," *Phys. Teach.* **33**, 503, 507–511 (Nov. 1995); Doug Huffman and Pat Heller, "What does the Force Concept Inventory actually measure?" *Phys. Teach.* **33**, 138–143 (March 1995).
4. Richard 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 (Jan. 1998).
5. See P. Heller, R. Keith, and S. Anderson, "Teaching problem solving through cooperative grouping. Part 1: Group versus individual problem solving," *Am. J. Phys.* **60**, 627–636 (July 1992), and P. Heller and M. Hollabaugh, "Teaching problem solving through cooperative grouping. Part 2: Designing problems and structuring groups," *Am. J. Phys.* **60**, 637–644 (July 1992).

6. The Force Concept Inventory was originally published by David Hestenes, Malcolm Wells, and Gregg Swackhamer, "Force Concept Inventory," *Phys. Teach.* **30**, 141–158 (March 1992). It was revised in 1995 by Ibrahim Halloun, Richard Hake, Eugene Mosca, and David Hestenes and is available online at <http://modeling.la.asu.edu/R&E/Research.html>.

7. The standard measurement error of a test is related to the reliability of the test and the standard deviation of the obtained test scores by:

$$\sigma_e = \sigma_t \sqrt{1 - \alpha},$$

where σ_e is the measurement error of the test, σ_t is the standard deviation of the distribution of obtained test scores, and α is the reliability of the test (measured with Chronbach's Alpha). For our students, σ_t is 5.17 items, α is 0.85, making the standard error of measurement 2.0 items. This means that the 95% confidence interval around a student's FCI score (2 standard errors of measurement) would be ± 4.0 items. Details are available in many statistics books, or see R.L. Thorndike and R.M. Thorndike, "Reliability" in *Educational Research, Methodology, and Measurement: An International Handbook*, 2nd ed., edited by John P. Keeves (Cambridge University Press, Cambridge, UK, 1997), pp. 775–790.

8. The standard error of measurement of the average FCI score for a class of 100 students would be given by:

$$\text{Standard Error} = \frac{\text{Standard Deviation}}{\sqrt{\text{Number of Students} - 1}} = \frac{5.71}{\sqrt{99}} = 0.520 \text{ FCI Items.}$$

9. Robert J. Beichner, "Testing student interpretation of kinematics graphs," *Am. J. Phys.* **62**, 750–762 (August 1994).

10. This supports the recommendation of the FCI authors — see Ref. 6.

Explorations in Physics

A Project-Centered Curriculum for Non-Scientists

DICKINSON COLLEGE, CARLISLE, PENNSYLVANIA

JUNE 15-20, 2003

Explorations in Physics (EiP): A modular curriculum developed as part of the ongoing Workshop Science Project. Intended for non-science majors, EiP units blend guided activities with capstone projects that connect physics to real world phenomena. Units on Motion, Light, Heat and Pressure have been published by John Wiley & Sons. New units on Sound, Energy, and Electricity are being tested.

WHO SHOULD PARTICIPATE:

Instructors at colleges and universities who teach physics to students whose interests lie outside the sciences, including pre-service teachers and students trying to satisfy a natural science requirement. There may be a limited number of spaces available for high school physics teachers as well.

Participants will receive hands-on experience with *Explorations in Physics* units that use computer tools for data collection, display, and analysis. Strategies will be discussed for managing workshop-style classes and mounting student-directed projects, as will the role of assessment and evaluation. Free tuition, room, board, and sample curricular materials will also be provided.

FOR APPLICATION MATERIALS:

Consult our web site at: <http://physics.dickinson.edu/EiP.html> or contact: Gail Oliver, Department of Physics and Astronomy, Dickinson College, Box 1773, Carlisle, PA 17013, O: 717/245-1845, F: 717/245-1642, Email: oliver@dickinson.edu

APPLICATION DEADLINE:

February 15, 2003

Funded by the Fund for the Improvement of Postsecondary Education Grant # P116B000927 as part of the continuing *Workshop Science* project.