Running Head: SIMPLIFIED FORCE CONCEPT INVENTORY

Can Assessment of Student Conceptions of Force be Enhanced Through Linguistic Simplification? A Rasch Model Common Person Equating of the FCI and the SFCI

Sharon E. Osborn Popp and Jane C. Jackson
Arizona State University

Paper presented at the annual meeting of the American Educational Research Association
San Diego, CA, April, 2009

Abstract

Assessment materials that are cognitively appropriate for younger students are needed as more high schools introduce physics education in ninth grade. In this paper, we report findings for the first of a series of studies investigating the validity of a simplified version of the Force Concept Inventory, a widely used measure of mechanics conceptual knowledge. A Rasch model common person equating was conducted to assess the comparability of the FCI and the simplified version (SFCI). Eleventh and twelfth grade physics students responded to both the FCI and the SFCI. Results provide evidence that the two instruments are measuring the same construct, at virtually the same level of difficulty. We also report preliminary results for a current study, which examines the responses of ninth graders and upperclassmen randomly assigned the test forms. Initial results are promising, suggesting that the SFCI provides an effective accommodation for students with a need for simplified language without providing an unfair advantage to students with no need. A viable simplified version of the FCI would be an important alternative for physics educators and physics education researchers. Benefits of a simplified FCI would include improved assessment for younger students as well as English language learners.

Can Assessment of Student Conceptions of Force be Enhanced Through Linguistic Simplification? A Rasch Model Common Person Equating of the FCI and the SFCI

In 2002, the American Association of Physics Teachers (AAPT) adopted a formal position promoting an approach to physics education called "Physics First." Physics First refers to the teaching of physics early in the high school program of study, in order to expose more students to physics and lay the foundation for all students to participate in more advanced science courses. AAPT has called for the development of materials appropriate to the cognitive development of students that will be "appreciably younger than students in traditional high school physics courses" (AAPT, 2002). The Force Concept Inventory (FCI), an assessment of conceptual understanding in Newtonian mechanics (Hestenes, Wells, & Swackhamer, 1992) has become the most widely used measure of mechanics concepts by physics educators and physics education researchers (see e.g., Hake, 1998 and Savinainen & Scott, 2002). The accumulation of evidence that either supports or challenges the validity of a simplified version of the FCI will be essential to physics educators and to the physics education research community, as the average age of students entering physics becomes lower and more suitable assessment tools are sought.

Linguistic Simplification

Assessment development guidelines routinely recommend careful attention to the reading level of non-language assessments, such as science, mathematics, and social studies tests, to avoid confounding the measurement of content with the measurement of reading ability. The challenges inherent in the valid assessment of students with limited proficiency in English provide a fitting analogy. If a student responding to a mathematics test cannot fully understand the language in the questions posed, her score will not provide the basis for an accurate inference regarding her mathematics ability. Similarly, if a student responding to the FCI cannot fully comprehend the complex wording or context in some items, his responses may not accurately reflect his conceptions of force. Research on English language learners has found linguistic simplification to be a viable assessment accommodation for students, particularly in the areas of mathematics and science (Abedi, 2006). One concern regarding the effects of accommodating students through linguistic simplification has been whether simplification provides an unfair advantage. While research in this area is limited, most studies have concluded that simplified English forms may be beneficial to all students and that no advantage is provided by the simplified form (e.g., Kiplinger, Haug, & Abedi, 2000). Rivera and Stansfield (2004) likened linguistically simplified test items to eyeglasses, i.e., improving vision for those in need, but not aiding those without impaired vision.

The Simplified Force Concept Inventory

Anecdotal reports by physics educators participating in physics education professional development workshops (e.g., the annual summer workshops of Modeling Instruction in Physics, Arizona State University) have indicated concern regarding the possibility that student FCI responses could be hindered by the complex wording and unfamiliar contexts presented in some items. The potential for younger (e.g., ninth grade) physics students to struggle with item complexity was of particular concern. A simplified version of the FCI (Jackson, 2007) has been developed to investigate whether a linguistically simplified version of the FCI could be a viable option for physics educators and physics education researchers. The items on the FCI were modified to fit a grade seven reading level and input regarding item revisions was incorporated

from the following sources: (a) several ninth grade physics educators, (b) the use of several item contexts from McCullough's gender-based revision (McCullough, 1994), and (c) suggestions from FCI co-author David Hestenes. Examples of the types of changes made to simplify items include replacing passive voice with active voice, replacement of unusual vocabulary words with higher frequency words, and rephrasing to eliminate conditional clauses. Many item contexts were replaced with more familiar contexts. For example, an item that referred to the positions of "two blocks at successive .20-second time intervals" was changed to refer to the positions of "two joggers at each second of time."

The focus of this paper is to describe an equating study conducted to compare the performance of physics students on the FCI with the performance of the same students on a simplified version of the FCI (SFCI). We also report preliminary findings from a current study investigating the viability of the SFCI for students at different grade levels. The following basic questions were addressed:

1. Was student performance comparable between the FCI and SFCI?
2. What differences, if any, were found in the item functioning of items modified for the SFCI?
3. Was there a difference in performance between the FCI and SFCI for students at different grade levels?

The first question, addressing comparability, requires the collection of evidence that either supports or opposes the contention that the FCI and SFCI measure the same construct. The second question investigates expected and unexpected differences in individual item functioning related to item modifications, and whether observed differences are invariant across age groups. Initial findings on item functioning are reported in this paper. The third question, regarding differences in performance, examines whether or not the SFCI affords an unfair advantage over the FCI. Even if the FCI and SFCI are measuring the same construct, the forms could be assessing students at different levels of difficulty. The SFCI was developed with the intention of accommodating the perceived need for simplified test language for younger students; performance is only expected to be significantly higher on the SFCI than the FCI for students with a need for simplified test language. If older students also perform significantly better on the SFCI than the FCI, then inferences based on raw scores regarding mechanics conceptual knowledge would not be considered equivalent between the two forms. Data collection is underway this spring for students of several physics teachers. Students of different age groups (ninth graders and upperclassmen) are being randomly assigned one of the two forms, within each classroom, for a post-mechanics curriculum test administration. Preliminary results for ninth grade data already submitted are reported.

Method

Data collection and analyses for the common person equating study will be described first, and then the data and analyses for the preliminary analyses conducted in the on-going study will be described.

Instruments

The instruments used in all studies discussed in this paper are the FCI (1995 version) and SFCI (2007). Both forms are available online to authorized educators and researchers at: http://modeling.asu.edu/R&E/Research.html. The Cronbach's alpha estimates of reliability for the FCI and SFCI responses were both .89.

Common Person Equating of FCI and SFCI

Sample (Spring 2008). High school physics students, in grades 11 and 12, from two different schools in two different states responded to the FCI after completing the mechanics curriculum in first year physics courses ($N = 95$). The same students were then administered the SFCI, between two to five months later.

Analyses. Basic item analyses were performed on the FCI and SFCI responses for the data collected in the common person equating study to evaluate basic item and test performance. A *t*-test for dependent samples was conducted to evaluate whether the difference in performance between the two forms was significant, employing an alpha level of .05.

A common person equating was conducted employing the Rasch (one-parameter logistic IRT) model for dichotomous items (Rasch, 1960/1980; Wright & Stone, 1979) to determine whether the performances of the students on both test forms are similar enough to support the contention that the same measurement construct is being assessed. Under the Rasch model, a correct response is modeled as a logistic function of the difference between an estimate of an examinee's ability and an item's difficulty. Estimates of examinee ability and item difficulty can be compared on the same linear logistic scale (in log-odd units, or logits). Positive logit values represent higher ability and higher degree of item challenge while negative logit values represent lower ability and lower degree of item challenge.

Student ability parameters were estimated for the FCI and SFCI responses and plotted against each other, with a 95% confidence band to examine the relationship between tests. The confidence band is constructed using the individual error estimates provided for each ability estimate and then plotting control lines around the ideal modeled relation between the two tests (i.e., the 45 degree identity line through the means). An additional analysis of the SFCI data was also conducted, anchoring student ability estimates to the ability parameters estimated in the FCI analysis, to obtain a comparable measure of average item difficulty between tests. The relationship between item difficulty estimates for the FCI and for the SFCI, anchored to ability estimates from the FCI analysis, was also plotted and examined. Infit and outfit mean squared fit statistics were also reviewed for items. Infit is a weighted mean square and outfit is an unweighted mean square residual that is sensitive to unexpected observations. No strict guidelines for interpretation exist, but many researchers look for values between 0.5 and 1.5, with 1.0 indicating best fit. Linacre and Wright (1994) have suggested a range of 0.7 to 1.3 as reasonable for non-high-stakes multiple choice tests. Fit values may be influenced by a small number of extreme observations or by sample size, so caution and common sense are often recommended in the interpretation of fit statistics (see e.g., Bond & Fox, 2001; Smith, 2000).

Preliminary Analyses Regarding Performance on FCI and SFCI

Sample (Spring 2009). Grade 9 first-year physics students ($N = 103$), from one high school, responded to either the FCI or SFCI after completing the mechanics curriculum, as randomly assigned by their teacher, i.e., distributing the forms in an alternating fashion within each section's administration. Fifty-one students responded to the FCI and fifty-two students responded to the SFCI, with a reasonable balance between genders in each group (i.e., 29 girls in the FCI group and 26 girls in the SFCI group).

Analyses. Basic item analyses were also performed on the FCI and SFCI responses for the early data collected in the on-going study to assess item and test functioning. A *t*-test for independent samples was conducted to evaluate whether the difference in performance between the two forms was significant, employing an alpha level of .05.

Results

Common Person Equating of FCI and SFCI

FCI and SFCI results were very similar for the same grades 11 and 12 students, with mean FCI and SFCI raw scores of 18.48 ($SD = 6.48$) and 18.42 ($SD = 6.64$), respectively. Summary statistics from conventional item analyses for the FCI and SFCI are presented in Table 1. The means were not significantly different from each other, with a t ($df = 94$) of .19, $p = .85$. The 95% confidence interval for the difference extends from -.60 to .72. The correlation between FCI and SFCI raw scores for the same grades 11 and 12 students was .88.

Table 1.

Summary Statistics from Item Analyses for FCI and SFCI[a] *for Grades 11 and 12 students*[b]

	FCI	**SFCI**
Mean	18.48	18.42
Standard Deviation	6.48	6.64
Median	17.00	17.00
Standard Error of Measurement	2.18	2.18
Mean Proportion Correct	.62	.61
Mean Item-Total Correlation	.47	.48

[a] 30 items on each form.

[b] $N = 95$; each student responded to both test forms.

Student ability parameter estimates also differed little for the Rasch analyses of the FCI and SFCI responses, with a mean student ability estimate of .81 logits ($SD = 1.47$) for the FCI and a mean student ability estimate of .86 logits ($SD = 1.48$). One extreme score (i.e., 100%)

within each set of responses (not the same case) was not included in the computation of the mean logit for each form; including the ability estimates in each computation of the mean would produce means of .85 (*SD* = 1.52) and .90 (*SD* = 1.54) logits for the FCI and SFCI, respectively. The correlation between FCI and SFCI student ability estimates was .89. Figure 1 shows a very strong relationship between FCI student ability estimates and SFCI student ability estimates, (adjusted by the mean difference of -.047) with only two out of 95 students falling clearly outside of the 95% confidence band.

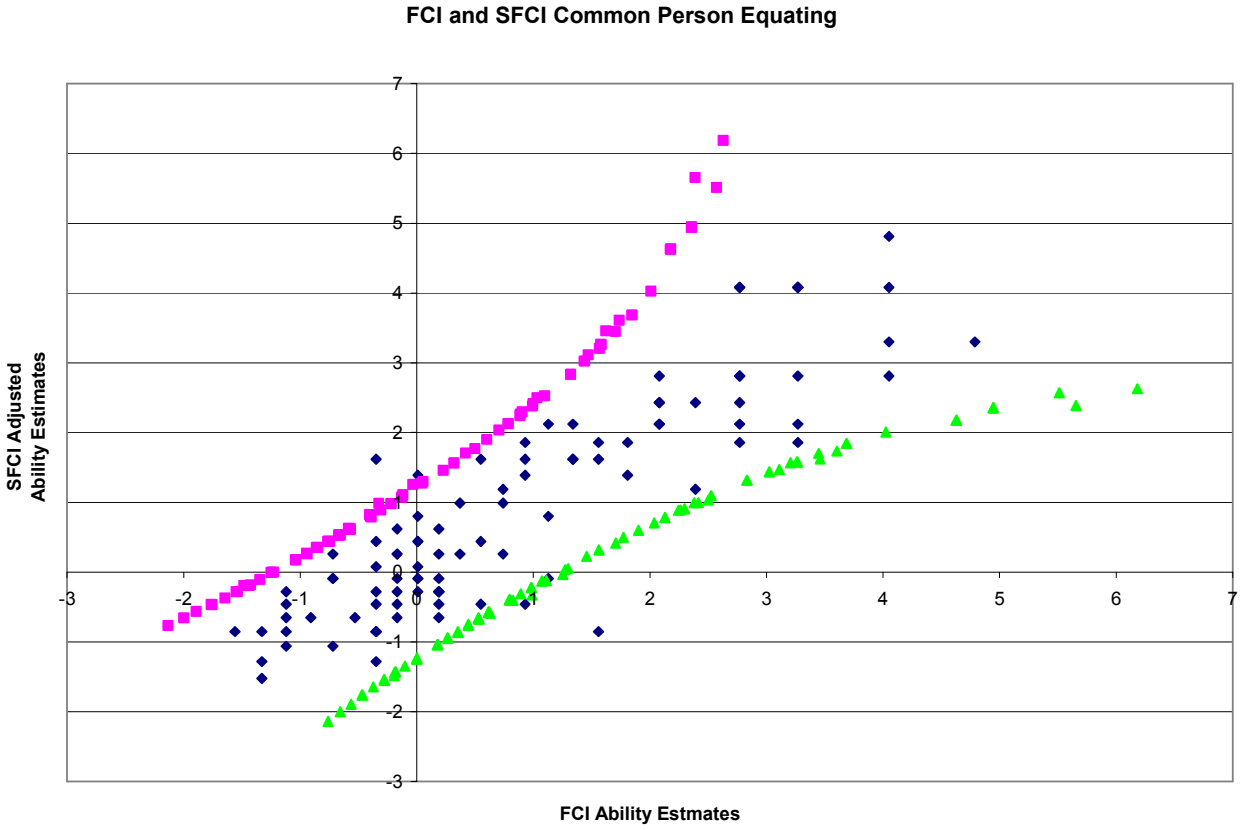

Figure 1. Rasch student ability parameter estimates for grades 11 and 12 students on the FCI and SFCI, with 95% confidence band.

Item difficulty parameter estimates also differed little between the Rasch analyses for the FCI and SFCI responses of grade 11 and 12 students. The average item difficulty estimate obtained in the anchored analysis of the SFCI was -.05 logits (*SD* = 1.43). The FCI analysis was fixed to mean 0 logits (*SD* = 1.29), resulting in a difference between the FCI and SFCI item difficulty estimates of .05 logits. The correlation between the two sets of difficulty estimates was .91. Figure 2 provides the plotted item difficulty estimates for the FCI and SFCI analyses. Most items except item 18 fall inside the 95% confidence band, with a few items (e.g., items 4, 21, and 25) falling close to the confidence bands. Item 1 is clearly the least challenging item on both forms and item 9 is the most difficult. For the FCI analysis, infit statistics ranged from .7 to 1.4

and outfit statistics for most items ranged from .5 to 1.3, with four values over 1.6, including a high outfit mean square value of 2.9 for item 4. For the SFCI anchored analysis, infit statistics ranged from .7 to 1.7 and outfit statistics for most items ranged from .6 to 1.7, with four values over 1.7, including items 4 and 21. Item difficulty estimates and fit statistics are provided in Table 2.

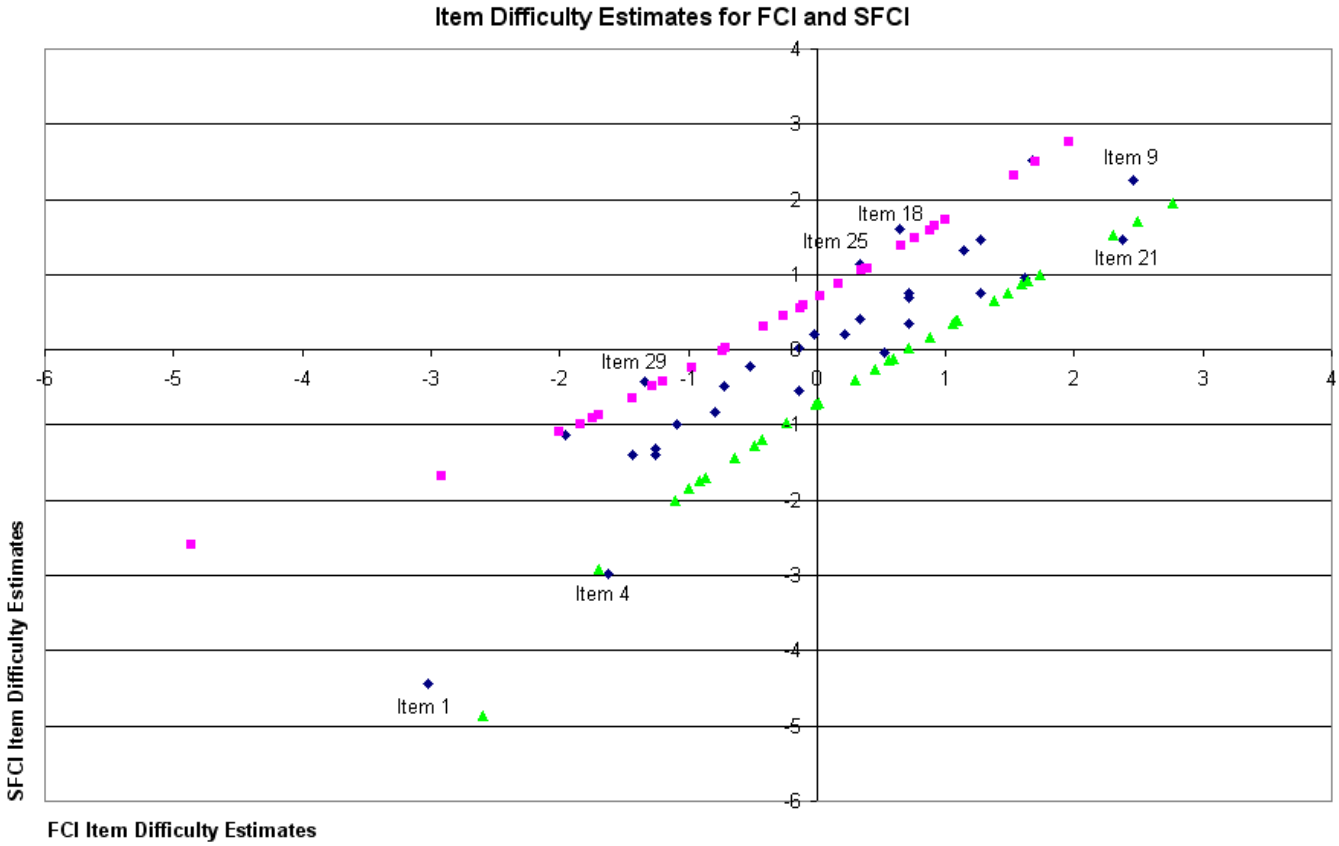

Figure 2. Rasch item difficulty parameter estimates for the FCI and SFCI, with 95% confidence band.

Table 2.

FCI and SFCI Item Difficulty Estimates with Mean Square and Standardized Fit Statistics

	FCI[a]						SFCI[b]					
Item	Difficulty Estimate (in logits)	Model SE	Infit MSQ	Infit STD	Outfit MSQ	Outfit STD	Difficulty Estimate (in logits)	Model SE	Infit MSQ	Infit STD	Outfit MSQ	Outfit STD
1	-3.02	0.52	1.0	-0.1	0.6	-0.4	-4.48	1.01	1.0	0.0	0.7	-0.1
2	-0.02	0.25	1.2	2.1	1.1	0.4	0.16	0.25	1.4	3.2	1.7	2.2
3	-0.52	0.26	0.8	-1.7	0.9	-0.4	-0.26	0.25	1.1	0.7	1.1	0.3
4	-1.62	0.32	1.0	0.2	2.9	2.1	-3.02	0.52	1.0	-0.1	3.7	1.2
5	1.27	0.26	0.8	-1.9	0.7	-2.0	1.40	0.26	1.0	-0.4	1.1	0.6
6	-1.96	0.35	1.1	0.4	0.8	-0.3	-1.17	0.29	1.1	0.5	0.9	-0.2
7	-0.72	0.26	1.2	1.3	1.2	0.5	-0.02	0.25	1.1	0.5	1.2	0.7
8	0.52	0.25	1.1	1.1	1.0	0.2	-0.08	0.25	1.1	0.9	1.1	0.2
9	2.46	0.29	1.2	1.1	1.3	0.8	2.21	0.28	1.3	1.8	1.7	2.0
10	-0.72	0.26	0.9	-0.7	0.7	-0.9	-0.52	0.26	1.1	0.5	2.3	2.6
11	0.71	0.25	0.7	-3.6	0.5	-3.0	0.64	0.25	1.3	2.1	1.5	2.0
12	-1.09	0.28	1.2	1.5	2.0	1.7	-1.01	0.28	1.0	-0.1	1.4	0.7
13	0.71	0.25	0.8	-2.2	0.7	-1.6	0.71	0.25	0.8	-2.2	0.6	-2.1
14	-0.14	0.25	1.1	0.7	1.2	0.6	-0.02	0.25	1.1	1.2	1.2	0.8
15	-1.25	0.29	1.1	0.5	1.9	1.4	-1.34	0.3	1.0	0.1	1.3	0.5
16	-1.43	0.30	1.0	-0.1	0.9	-0.3	-1.43	0.3	1.1	0.4	0.9	-0.1
17	0.34	0.24	1.2	1.4	1.1	0.4	0.34	0.24	1.1	1.1	1.1	0.5
18	0.64	0.25	0.9	-0.8	0.9	-0.6	1.54	0.26	0.7	-2.5	0.6	-2.0
19	0.71	0.25	1.0	0.0	1.0	-0.2	0.28	0.24	0.9	-1.5	0.7	-1.4
20	-0.14	0.25	1.1	0.8	0.9	-0.5	-0.59	0.26	1.1	0.7	0.9	-0.3
21	2.37	0.29	1.4	2.3	1.6	1.7	1.40	0.26	1.7	4.4	2.4	4.9
22	0.22	0.24	1.0	-0.2	1.1	0.4	0.16	0.25	1.0	0.4	1.1	0.5
23	1.14	0.25	1.1	0.8	1.0	0.2	1.27	0.26	1.1	0.7	1.1	0.7
24	-0.79	0.27	0.8	-1.6	0.6	-1.1	-0.86	0.27	1.0	-0.2	0.8	-0.5
25	0.34	0.24	0.9	-0.6	0.9	-0.6	1.08	0.25	1.1	0.7	1.1	0.5
26	1.68	0.27	0.9	-0.8	0.9	-0.4	2.45	0.29	0.9	-0.4	0.8	-0.6
27	1.27	0.26	1.0	-0.4	0.8	-1.0	0.71	0.25	1.0	0.3	0.9	-0.4
28	-1.25	0.29	0.9	-0.7	0.6	-0.9	-1.43	0.3	1.0	-0.2	1.4	0.6
29	-1.34	0.30	1.0	-0.3	1.2	0.4	-0.45	0.25	1.2	1.9	4.2	5.3
30	1.61	0.26	0.8	-1.4	0.8	-1.1	0.89	0.25	1.0	0.1	0.9	-0.4

[a]FCI: RMSE = .28; Adj. *SD* = 1.26; Separation = 4.46; Reliability = .95

[b]SFCI: RMSE = .33; Adj. *SD* = 1.40; Separation = 4.27; Reliability = .95

Item analyses, Rasch item-difficulty estimates, and Rasch fit statistics were examined to explore how modified items functioned differently between the original and simplified forms of the FCI. The items expected to vary most were items 17, 25, and 29, which had received the highest degree of modification to the response options. Items 25 and 29 did vary to some degree, but item 17 did not, with the same proportion correct, difficulty estimates, and very similar patterns of option response between the two forms. Item 25 was slightly more difficult on the SFCI, but otherwise showed similar item functioning in pattern of option response. Item 29 was also more difficult on the SFCI, and had the highest outfit mean square fit value. Feedback revealed that the revised context (a motionless fish) increased the complexity of the item. Item 18 was more difficult on the SFCI, fell outside the confidence band in the comparison between forms, and elicited feedback that the revised item on the SFCI was "wordier" than the original item. Item 4 was less challenging on the SFCI, but was one of the least challenging items on both forms. Item 21 was less challenging on the SFCI, but was one of the most difficult items on both forms. Item 18 was already revised again prior to the study currently underway, and revisions to other items will be considered following complete results available later this spring.

Preliminary Analyses Regarding Performance on FCI and SFCI

Data collection is underway this spring for ninth grade and upperclassmen physics students. Students have, or will be, randomly assigned either the FCI or SFCI in posttest administrations after completing the mechanics curriculum. Data from some ninth grade students has been analyzed and preliminary results are presented here.

Summary statistics from item analyses for the ninth grade students that took either the FCI or SFCI are reported in Table 3, including mean raw scores for each form. The mean FCI raw score ($N = 51$) was 13.10 ($SD = 4.30$) and the mean SFCI raw score ($N = 52$) was 14.77 ($SD = 3.89$). The students performed significantly higher on the SFCI with a t ($df = 101$) of -2.07, $p = .04$. The 95% confidence interval for the difference extends from -3.27 to -.07.

Table 3.
Summary Statistics from Item Analyses for FCI and SFCI[a] *for Grade9 students*[b]

	FCI	**SFCI**
Mean	13.10	14.77
Standard Deviation	4.30	3.89
Median	13.00	15.00
Standard Error of Measurement	2.16	2.31
Mean Proportion Correct	.44	.49
Mean Item-Total Correlation	.35	.30

[a] 30 items on each form.

[b] $N = 103$; students were randomly assigned to forms with $N = 51$ for FCI and $N = 52$ for SFCI.

Discussion

Grades 11 and 12 students performed similarly on the FCI and on a simplified form of the FCI, taken months later. No significant difference was found between mean test scores. The results of a Rasch common person equating study indicate little difference in student performance between the two instruments. The common person equating results provide strong evidence that the two instruments are measuring the same construct, at virtually the same level of difficulty, indicating that students did not receive an unfair advantage on the simplified version of the FCI. Initial evidence also suggests that the most items on the FCI and SFCI are strongly related. A few items displayed differences in item functioning and were identified as requiring additional revision. Most revisions will not take place until a larger study examining FCI and SFCI results is completed, allowing for cross-validation of empirical findings and additional feedback from students and teachers. One limitation of the comparability study is that all students took the FCI prior to the SFCI, with a possible threat to validity that prior exposure to the FCI affected response on the SFCI in some systematic way. The equating study provided strong evidence that the SFCI can produce results comparable to the FCI. However, additional data that controls for prior exposure will be needed to confirm or refute the finding that the SFCI does not provide an unfair advantage over the FCI when administered to older students.

Research is also needed to determine whether the SFCI can effectively accommodate younger students by enhancing their ability to understand and respond to items through reducing item complexity. Preliminary findings for the ninth grade student data indicate that younger students perform significantly higher on the SFCI than on the FCI. A study currently underway is collecting additional data from ninth graders and upperclassmen, who have been, or will be, randomly assigned to respond to one of the forms, this spring.

The SFCI has the potential to be a valuable tool for physics educators and researchers who seek assessment materials that are more cognitively appropriate for younger students, especially as more high schools offer physics to ninth graders. A simplified version of the FCI may also improve the validity of assessment of physics students who are English language learners. Results of this study provide evidence that the FCI and SFCI instruments are measuring the same construct, at virtually the same level of difficulty for older students. Additional research is being conducted to further examine whether the SFCI can effectively accommodate students with a need for simplified language without providing an unfair advantage to students with no need. A viable simplified version of the FCI would be an important alternative for physics educators and physics education researchers, particularly for the assessment of younger students.

References

Abedi, J. (2006). Psychometric issues in the ELL assessment and special education eligibility. *Teachers College Record, 108*(11), 2282-2303.

American Association of Physics Teachers (AAPT). (2002). AAPT statement on Physics First. Retrieved July 23, from http://www.aapt.org/Policy/physicsfirst.cfm.

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. Mahwah, NJ: Lawrence Erlbaum Associates.

Hake (1998). Interactive-engagement vs. traditional methods: A six-thousand student survey of mechanics test data for introductory physics courses. *American Journal of Physics, 66*, 64-74.

Halloun, I., Hake, R., Mosca, E. & Hestenes, D. (1995). Force Concept Inventory (revised 1995). Retrieved July 9, 2008 from http://modeling.asu.edu/R&E/Research.html.

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force Concept Inventory. *Physics Teacher, 30,* 141-58.

Jackson, J. C. (2007). Force Concept Inventory (simplified). Retrieved June 25, 2008 from http://modeling.asu.edu/MNS/MNS.html.

Kiplinger, V. L., Haug, C. A., & Abedi, J. (2000, April). Measuring math – not reading – on a math assessment: A language accommodations study of English language learners and other special populations. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Linacre, J M., & Wright, B. D. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*, 370. Retrieved June 21, 2008 from http://www.rasch.org/rmt/rmt83b.htm.

McCullough, L. (1994). Gender, context, and physics assessment. *Journal of International Women's Studies, 5*(4), 20-30.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research, 1960. Expanded edition, Chicago: The University of Chicago Press, 1980.

Rivera, C., & Stansfield, C. W. (2004). The effect of linguistic simplification of science test items on score comparability. *Educational Assessment, 9*(3&4), 79-105.

Savinainen, A., & Scott, P. (2002). The Force Concept Inventory: A tool for monitoring student learning. *Physics Education, 37*(1), 45-52.

Smith, R. M. (2000). Fit analysis in latent trait measurement models. *Journal of Applied Measurement, 1*, 199-218.

Wright, B. D., & Stone, M. H. (1979). *Best Test Design: Rasch Measurement.* Chicago: MESA Press.