Paper Title Is the Force Concept Inventory Biased? Investigating Differential Item Functioning on a Test of Conceptual Learning in Physics

Author(s) Sharon E. Osborn Popp, Arizona State University; David Meltzer, Arizona State University; M. Colleen Megowan-Romanowicz, Arizona State University

Session Title Diversity and Bias

Session Type Paper

Presentation Date 4/9/2011

Presentation Location New Orleans, Louisiana, USA

Descriptors RASCH Models, Science Education, Validity/Reliability

Methodology Quantitative

Unit SIG-Science Teaching and Learning

Is the Force Concept Inventory Biased?

Investigating Differential Item Functioning on a Test of Conceptual Learning in Physics

Sharon E. Osborn Popp, David E. Meltzer, and Colleen Megowan-Romanowicz

Arizona State University

Abstract

Persistent differences in performance between females and males on measures of physics conceptual learning have prompted interest in investigating and reducing the gender gap. Educators and researchers need to have confidence in their interpretations of results and want to know if observed group differences are artifacts of test bias or due to factors like background or instruction. A *differential item functioning* (DIF) analysis was conducted on responses to a widely used measure of conceptual learning to assess whether properties of the test itself, unrelated to student ability, influence performance by gender. Findings provide evidence that the test is not systematically biased in favor of males. However, three items did exhibit substantial DIF, two favoring males and one favoring females.

Is the Force Concept Inventory Biased? Investigating Differential Item Functioning on a Test of Conceptual Learning in Physics

Persistent differences in performance between females and males on measures of physics conceptual learning have led to substantial interest in investigating and reducing the gender gap. One of the most frequently used measures of conceptual learning in force and related kinematics is the Force Concept Inventory (FCI) (Hestenes, Wells, & Swackhamer, 1992). Concerns suggesting that the disparity between female and male performance may be due to properties of test items unrelated to the measurement construct (i.e., that situational contexts of some items may be more familiar to males, and thus more favorable) have led to increased awareness of the potential for item-bias on assessments like the FCI. However, differences in item performance may not necessarily reflect systematic bias attributable to the measurement instrument. Differences in performance could reflect actual differences between groups (e.g., due to differences in background or opportunity to learn). Educators and researchers need to have confidence in their interpretations of results and want to know if observed group differences are artifacts of test bias or due to factors like background or instruction. In this study, a *differential item functioning* (DIF) analysis was conducted to detect patterns of item response that can provide evidence as to whether or not FCI results may be substantially influenced by systematic item-bias.

Background

The Gender Gap

Females continue to score significantly lower than males on science achievement tests, with the greatest gender disparity in physics (Kahle & Meece, 1994; Mullis, Martin, Fierros, Goldberg, & Stemler, 2000). Persistent female-male disparity has been observed in FCI results

and attempts to explain the gender gap via background variables or reduce the gap via instruction have shown mixed results thus far. Lorenzo, Crouch, and Mazur (2006) found that the use of interactive engagement instructional methods (i.e., that emphasize in-class interaction and cooperative problem-solving activities) could reduce or eliminate the gender performance gap, as measured by the FCI, in college introductory physics courses. Pollock, Finkelstein, and Kost (2007), using interactive engagement methods similar to Lorenzo et al., found that the gap persisted on posttest measures using another conceptual learning inventory (Thornton & Sokoloff's Force and Motion Conceptual Evaluation, 1998), and followed up by investigating background variables such as prior related knowledge and attitudes (Kost, Pollock, & Finkelstein, 2009). Docktor and Heller (2008) found substantial gender differences on FCI performance, despite little or no difference between females and males on course or final exam grades.

Other research has focused on the nature of the FCI itself, and examined whether the instrument could be biased in favor of males. McCullough (2004) noted that the situational contexts of FCI items are male-oriented or lab-oriented (e.g., rockets, cannons, and steel balls) and developed a female-centric version of the FCI which maintained the physics content of the items but "used *stereotypically* female contexts such as shopping, cooking, jewelry, and stuffed animals" (p. 24). Males performed less well on the female-context FCI than on the original, demonstrating that context can affect performance. However, females did not perform significantly better on the female-context version than on the original. McCullough did note that some items had different patterns of response by gender between the two versions. In a different study that compared results between the original and female-context versions of the FCI, McCullough and Meltzer (2001) identified specific items with notable gender differences in

response. Females showed a much higher rate of correct response on the female-context version for FCI items 14 and 23. Items 14 and 23 were also found to have the largest female-male differences in correct response by Docktor and Heller (2008).

Differential Item Functioning

We expect students at different ability levels to perform differently on test items. We do not, however, expect examinees that are comparable with respect to their level of ability or learning to perform at substantially different levels on the same test items. The Standards for Educational and Psychological Testing (AERA, APA, & NCME, 1999), under the section on *Fairness in Testing and Test Use* explain that, "Differential item functioning exists when examinees of equal ability differ, on average, according to their group membership in their particular responses to an item." (p. 81). Statistical methods of investigating *differential item functioning* (DIF) have evolved considerably in recent decades and have become a standard part of large-scale assessment programs (see e.g., Camilli, 2006; Holland & Wainer, 1993).

DIF analyses detect unexpected differences in item response between groups of examinees that are expected to respond similarly because they are at the same level of ability. Depending on the type of statistical analysis, the proxy for level of ability may be observed test score or an estimate of ability calculated via latent-trait methods like IRT and the Rasch model. If there is an unexpected difference in response between examinees from the same population that are at the same ability level but belong to different identifiable groups, the item is said to exhibit DIF. The presence of DIF does not necessarily mean that an item is biased, however. The DIF finding may be due to chance and may not be replicated in other samples. The DIF finding may be due to factors outside the testing situation that interact with the examinee's consideration of the item's content, such as differential background preparation. An item that displays DIF

must be reviewed judgmentally to determine whether properties of the item (e.g., wording or context not directly related to the content being measured) may be responsible for the unexpected differential performance, and if so, whether revision or deletion is recommended.

The purpose of this paper is to describe a DIF analysis conducted to investigate the presence of unlikely differences in individual FCI item performance based on gender. Posttest FCI responses of 4775 high school students from first-year physics classes were analyzed. Results can assist physics education researchers and physics educators in understanding whether FCI results are a valid estimate of student conceptual understanding or if results may be substantially influenced by construct-irrelevant variance due to systematic item-bias.

Method

Sample

FCI student responses were collected from 95 high school physics teachers around the United States during four academic years in the late 1990s. The physics teachers provided student data in the course of participation in National Science Foundation sponsored workshops in the modeling method of physics instruction (see http://modeling.asu.edu/index.html). Students responded to the FCI after completing the mechanics curriculum in first year (regular and honors) physics courses ($N = 4775$). The responses contain 49% female responses ($N = 2348$) and 51% male responses ($N = 2427$).

Instrument

The FCI is a 30-item multiple-choice instrument developed to assess student understanding of basic concepts of Newtonian mechanics (Hestenes, Wells, & Swackhamer, 1992). Revised in 1995 (Halloun, Hake, Mosca, & Hestenes, 1995), and translated into eighteen languages, the FCI 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 FCI (1995 revised version) is available online to authorized educators and researchers at: http://modeling.asu.edu/R&E/Research.html.

Analyses

Differential item functioning was examined using the Rasch (one-parameter logistic IRT) model for dichotomous items (Rasch, 1960/1980; Wright & Stone, 1979). 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. The probability (P) of a correct response, given ability (b) and difficulty (d) is given by:

$$P(b, d) = \frac{e^{(b-d)}}{1 + e^{(b-d)}} \quad (1)$$

Where: $e = 2.718$ (base of the natural log system)
b = student ability
d = item 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 item difficulty while negative logit values represent lower ability and lower item difficulty.

The Rasch model requires that item difficulty estimates be invariant across samples of examinees from the same population. Items that vary substantially between identified groups are said to exhibit DIF and should be investigated. The difference between each item's estimated difficulty parameter (in logits) for each gender group is called the DIF contrast. The statistical significance of the difference can be assessed with a t-test based on the joint standard error for the two estimates, for each item. A Bonferroni adjustment for the probability values of the significance tests on 30 items was applied for a type I error rate of 0.05, resulting in a criterion of

0.05/30 = 0.0017. Even with a low probability threshold, sample size can influence the statistical detection of very small amounts of DIF; therefore assessing practical significance becomes necessary. Camilli (2006) cautions that "inferential test statistics are not appropriate as measures of the practical size of DIF, and they should not be used as effect sizes" (p. 240). Wang (2009) explains that for Rasch models, the DIF contrast value in logits is the appropriate effect size value, as the "DIF amount of *d* logits represents an odds-ratio of 2.72^{d}" (p. 107). Thus, a DIF amount of 0.5 logits corresponds to an odds-ratio of 1.65, which is often used as a cut-off point for substantial DIF. Test statistics will be reported, but a threshold of 0.5 logits will be applied in determining items with meaningful DIF.

Point-biserial correlations and item mean squared fit statistics (infit and outfit) were reviewed to assess model fit and unidimensionality. 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 of fit statistics 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. Descriptive statistics, estimates of reliability, proportion correct values, and correlations between observed proportions of correct response and between parameter estimates were also computed and reviewed. Distribution maps of student-ability parameter estimates and item-difficulty parameter estimates are presented for comparison between analyses of females and males. Graphs that plot the item characteristic curves representing the probability of correct response as a function of person ability (for each gender) are provided for items showing substantial DIF.

Results

The Cronbach's alpha estimates of internal consistency reliability for the FCI were 0.88 for the total sample, 0.84 for females only, and 0.89 for males only. The Rasch person separation reliability estimates were 0.86 for the total sample, 0.84 for females only, and 0.86 for males only. Item point-biserial correlations ranged from 0.29 to 0.62 (M = 0.46; SD = 0.08) for the total sample, 0.19 to 0.60 (M = 0.42; SD = 0.10) for females only, and 0.31 to 0.66 (M = 0.48; SD = 0.09) for males only. For the total sample, infit mean square fit statistic values ranged from 0.79 to 1.23 and outfit mean square values ranged from 0.75 to 1.69. The highest outfit values (>1.30) were associated with items 21 and 29. For females only, infit values ranged from 0.82 to 1.28 and outfit values ranged from 0.77 to 1.49; items 21 and 22 had the highest outfit values. For males only, infit values ranged from 0.79 to 1.21 and outfit values ranged from 0.72 to 1.83; items 29 and 15 had the highest outfit values.

The mean FCI raw score for all students was 15.63 (SD = 6.74), with a mean proportion of correct response of 0.52 across the 30 items and a standard error of measurement (SEM) of 2.32. For females only, the mean raw score was 13.52 (SD = 5.96), with a mean proportion correct of 0.45 and SEM of 2.36. For males only, the mean raw score was 17.66 (SD = 6.74), with a mean proportion correct of 0.59 and SEM of 2.27. Males had a higher proportion of correct response on all 30 items, with raw score differences ranging from 0.03 to 0.28 (M = 0.14; SD = 0.06). Proportion correct values for each item by gender, ordered by raw score difference, are provided in Table 1. The correlation between proportion correct values for females and males was 0.89.

Table 1.

Proportion Correct Response Values on FCI Test Items for Females and Males, Ordered by Raw Score Difference

FCI Item	Proportion Correct Female	Male	Difference (M – F)
29	0.73	0.76	0.03
15	0.48	0.52	0.04
4	0.63	0.67	0.04
9	0.35	0.42	0.07
28	0.57	0.65	0.09
1	0.76	0.86	0.10
17	0.33	0.44	0.11
2	0.42	0.53	0.11
16	0.70	0.81	0.11
26	0.14	0.25	0.11
7	0.67	0.78	0.12
20	0.44	0.57	0.12
18	0.33	0.46	0.14
25	0.30	0.44	0.14
11	0.36	0.50	0.14
6	0.73	0.87	0.14
21	0.29	0.43	0.14
5	0.23	0.38	0.15
8	0.47	0.63	0.15
24	0.64	0.80	0.16
19	0.39	0.54	0.16
30	0.30	0.46	0.16
12	0.63	0.78	0.16
3	0.45	0.61	0.17
22	0.31	0.49	0.18
10	0.57	0.75	0.18
13	0.33	0.52	0.19
27	0.43	0.64	0.22

14	0.31	0.57	0.26
23	0.24	0.52	0.28

The mean person ability parameter estimate for all students in the Rasch analysis was 0.19 logits (*SD* = 1.37). For females, the mean ability parameter estimate was -0.24 (*SD* = 1.13); for males, the mean was 0.61 (*SD* = 1.49). The correlation between item difficulty parameter estimates for females and males was 0.89. The values of the DIF contrasts (i.e., differences between item difficulty parameter estimates for males and females) ranged from 0.0 to 0.73 logits (absolute value). Fourteen items had DIF contrasts with significant *t* statistics. Seven favored males (positive contrast values) and seven favored females (negative contrast values). Three items (23, 15, and 14) had contrasts exceeding 0.50 logits. Two favored males (23 and 14) and one favored females (15). Three other items (4, 29, and 9) had DIF contrast values close to the 0.50 logits cut-off (all favored females). Table 2 provides item parameter estimates and DIF contrast values, ordered by size of DIF contrast.

Table 2.

Rasch Item Difficulty Parameter Estimates, DIF Contrast Values, Standard Errors, and t Statistics, Ordered by Size of DIF Contrast Value

	Item Difficulty Parameter Estimates				DIF Contrast		
FCI Item	Female	SE	Male	SE	M – F	Joint SE	*t* value
30	0.76	0.05	0.76	0.05	0.00	0.07	0.00
19	0.30	0.05	0.30	0.05	0.00	0.07	0.00
8	-0.15	0.05	-0.15	0.05	0.00	0.07	0.00
7	-1.12	0.05	-1.12	0.05	0.00	0.07	0.00
5	1.21	0.05	1.21	0.05	0.00	0.07	0.00

16	-1.35	0.05	-1.37	0.06	0.02	0.08	0.32
21	0.87	0.05	0.92	0.05	-0.05	0.07	-0.70
25	0.81	0.05	0.87	0.05	-0.06	0.07	-0.88
11	0.44	0.05	0.51	0.05	-0.07	0.07	-1.00
3	-0.01	0.05	-0.09	0.05	0.08	0.07	1.22
26	1.97	0.07	2.06	0.06	-0.09	0.09	-1.01
18	0.62	0.05	0.73	0.05	-0.11	0.07	-1.59
1	-1.68	0.05	-1.82	0.06	0.14	0.08	1.73
22	0.73	0.05	0.59	0.05	0.15	0.07	2.14
20	0.00	0.05	0.17	0.05	-0.16	0.07	-2.43
13	0.62	0.05	0.42	0.05	0.20	0.07	2.96
12	-0.91	0.05	-1.16	0.05	0.25	0.07	3.49*
2	0.13	0.05	0.39	0.05	-0.26	0.07	-3.89*
24	-1.00	0.05	-1.27	0.06	0.27	0.07	3.70*
28	-0.60	0.05	-0.32	0.05	-0.28	0.07	-4.15*
17	0.60	0.05	0.88	0.05	-0.28	0.07	-4.08*
10	-0.62	0.05	-0.95	0.05	0.33	0.07	4.68*
27	0.08	0.05	-0.27	0.05	0.35	0.07	5.26*
6	-1.49	0.05	-1.92	0.07	0.44	0.08	5.27*
9	0.49	0.05	0.96	0.05	-0.47	0.07	-6.83*
29	-1.48	0.05	-0.99	0.05	-0.49	0.07	-6.72*
4	-0.92	0.05	-0.42	0.05	-0.50	0.07	-7.36*
14	0.72	0.05	0.15	0.05	0.57**	0.07	8.33*
15	-0.18	0.05	0.41	0.05	-0.59**	0.07	-8.94*
23	1.14	0.05	0.40	0.05	0.73**	0.07	10.30*

* $p < 0.0017$ ** Absolute value of DIF Contrast > 0.50 logits

A practical feature of the Rasch model is that person ability parameter estimates and item difficulty parameter estimates can be compared along the same logit scale. A vertical ruler, also called a construct map or Wright map (Wilson, 2005), plots the relative distributions of person

ability estimates and item difficulty estimates on the logit scale. Figure 1 contains Wright maps from the female and male Rasch analyses. Person ability estimates are distributed along the left side of each map, with positive logit values corresponding to higher estimated ability at the top, and negative logit values reflecting lower estimated ability at the bottom. Person abilities on the Wright maps show fewer students in the high-ability range for females compared to males. Item difficulty estimates are distributed in a similar manner on the right of each map, with more challenging items located toward the top and less challenging items toward the bottom. The relative position of most items is similar across the maps, except for items noted above that show higher DIF contrasts. For example, item 23, with a high positive contrast value, is lower (i.e., less challenging) on the male Wright map, while item 15, with a high negative contrast value, is lower on the female Wright map.

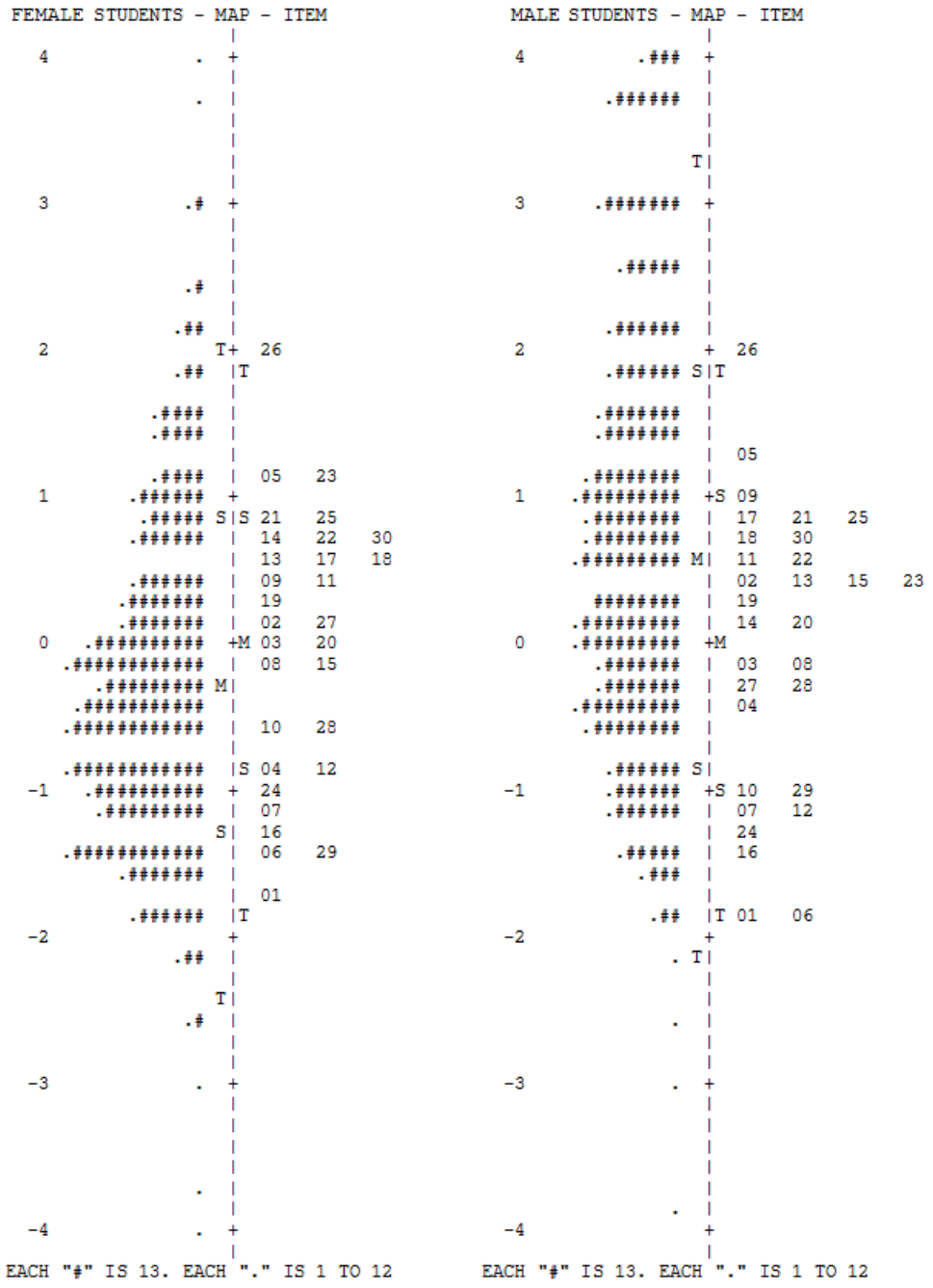

Figure 1. Wright maps of student ability estimates and FCI item difficulty estimates for females and males, respectively. Scale values are logits, with positive values (top) indicating higher ability/more difficulty and negative values indicating lower ability/less difficulty. Each "#" represents 13 students; each "." represents 1 to 12 students. M is mean, S is one sample standard

deviation from mean, T is two standard deviations from mean. Note that items that fall within the same bin are printed in sequential order along the same row.

Item characteristic curves for items 23, 15, and 14 are plotted in Figures 2, 3, and 4, respectively. Each graph shows, for each gender, the probability of responding correctly to that item for varying levels of ability. For example, in Figure 2, a female with estimated ability of 1.0 logits has a 0.50 probability of responding correctly to item 23, while a male with the same estimated ability of 1.0 logits has a 0.65 probability of responding correctly. In Figure 3, the situation is reversed; a female (ability = 1.0 logits) would have a 0.77 probability of responding correctly to item 15, while a male at the same ability would have a 0.65 probability of responding correctly.

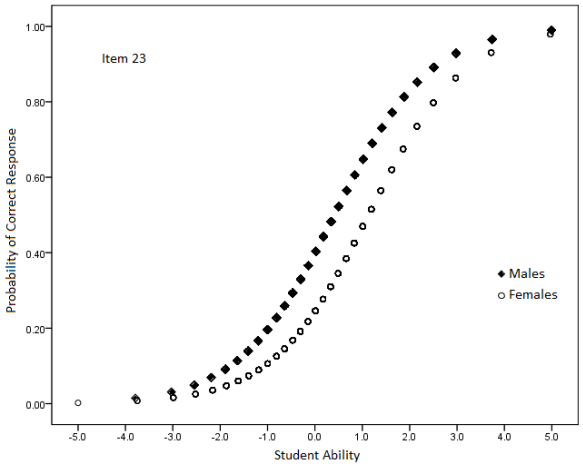

Figure 2. Item characteristic curves for males and females on Item 23.

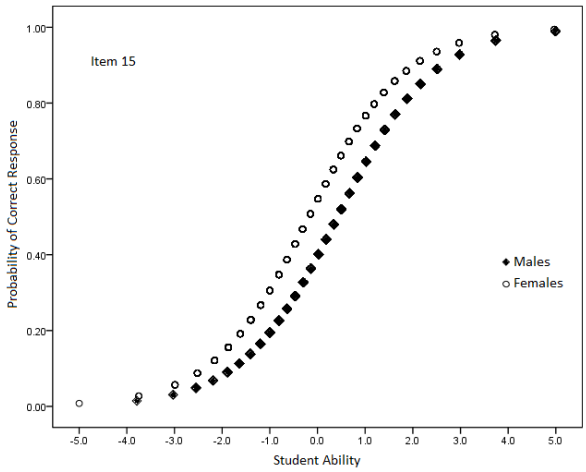

Figure 3. Item characteristic curves for males and females on Item 15.

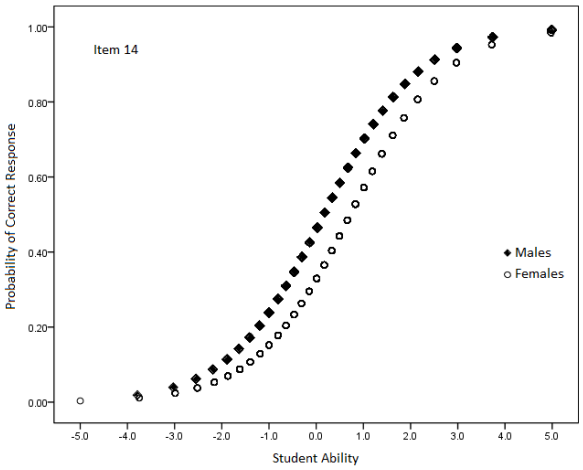

Figure 4. Item characteristic curves for males and females on Item 14.

Discussion

Results of the DIF analysis of the FCI indicate no clear trend in favor of males. Five items showed no DIF. Twelve items favored males, with seven DIF contrasts over 0.25 logits, and only two showed substantial DIF (over 0.50 logits). Thirteen items favored females, with seven DIF contrasts over 0.25 logits, and one showed substantial DIF (over 0.50 logits). The items exhibiting substantial DIF were 23, 15, and 14, with items 23 and 14 favoring males and item 15 favoring females. Of the three items exhibiting substantial DIF (items 23, 15, and 14), two of these items (23 and 14) have been previously cited in physics education research literature

as having the largest differences between females and males (Docktor and Heller, 2008) and when re-written with a female context, showed a much higher rate of correct responses by females (McCullough and Meltzer, 2001). A review of the wording and figures associated with items 14, 15, and 23, reveals that items 14 and 23 (that exhibited substantial DIF in favor of males) require predicting the path that a moving object will take from pictorial response options (a bowling bowl falling from an airplane and a rocket moving through space). Item 15 (that exhibited substantial DIF in favor of females) involves a compact car pushing a truck and requires choosing from fairly lengthy options that might be called "wordy." However, inspection of other items with similar response features and wording does not suggest a pattern of male advantage for items with illustrated path prediction options or female advantage on items that are heavily dependent on reading many words.

Review of the three items exhibiting DIF also indicated that one of the items (14) "stood alone" regarding the item context, while the others shared their context with at least one other item (15 is the first of two related items; 23 is the third of four related items). Responses to items that share a context are not necessarily highly dependent, but an unusually high degree of dependence between particular items can confound interpretations regarding examinee ability on the measurement construct. Local item dependence, where one's response to one item depends on one's response to another item, is a concern under most latent-trait models and has led to approaches that combine dependent items and treat them as superitems or testlets (see e.g., Wainer, Sireci, & Thissen, 1991; Wilson & Iventosch, 1988). Interrelationships between and among items that share a common context on the FCI may be worthy of further attention and research, since response patterns to highly interrelated items may provide additional insights into differential performance based on gender.

In addition to investigating the FCI for items that may be highly interrelated and examining whether assessing results under a different structure (e.g., as if some item clusters are superitems), there is also a need to conduct DIF analyses on other samples to confirm the present findings. Students in the sample used for this study were taught by teachers that had participated in national workshops that provided training in the modeling method of physics instruction. Length of workshop participation and degree of modeling instruction implementation varied considerably across teachers, so investigating whether degree of interactive engagement practices is related to DIF findings may also be worthwhile. Examining pretest FCI data, breaking down the current sample into regular and honors samples, and conducting a DIF analysis on FCI responses for second year high school physics students are also warranted.

Despite no obvious indication of bias upon review of the items exhibiting substantial DIF, educators and researchers may still be wary of the effect that the three items may have on score interpretations. If all three items flagged for substantial DIF on the FCI were removed from scoring, the female mean would be 12.49 (SD = 5.44) and the male mean would be 16.05 (SD = 6.17). If only item 23, the item with the highest DIF contrast value, were removed from scoring, the female mean would be 13.28 (SD = 5.82) and the male mean would be 17.14 (SD = 6.59). Average performance by females and males for raw scores, percentage correct scores, and the effect sizes for differences between genders, were found to differ little when the items are removed (Table 3). The effect size for the difference between male and female mean scores on the complete 30 items of the FCI is moderately high, with a Cohen's d value of 0.65. When scored without item 23, the effect size for the difference, though very slightly lower, remains moderately high, with a value of 0.62. With all three items exhibiting substantial DIF removed (items 23 and 14, which favored males and item 15, which favored females) the effect size for

the difference between males and females is also moderately high at 0.61. Removing the items that demonstrated substantial DIF does not change the inference that, on average, males perform significantly better than females on the Force Concept Inventory, after their first high school physics course. The effect is moderately large, regardless of scoring without high DIF contrast items.

Table 3.

Raw Score and Percentage Score Means, Standard Deviations, Male-Female Differences and Effect Sizes for Scoring FCI with and without high DIF-contrast items

# of Items Scored	Description	Gender	Raw Score Mean(*SD*)	Percentage Mean(*SD*)	M-F Percentage Difference	Effect Size for Difference*
30	All items	Female	13.52(*5.96*)	45.07(*19.87*)	13.80	0.65
		Male	17.66(*6.74*)	58.87(*22.47*)		
29	No 23	Female	13.28(*5.82*)	45.79(*20.07*)	13.19	0.62
		Male	17.14(*6.59*)	59.10(*22.72*)		
27	No 23, 15, 14	Female	12.49(*5.44*)	46.26(*20.15*)	13.31	0.61
		Male	16.05(*6.17*)	59.44(*22.85*)		

* Cohen's *d*

Educators and researchers need to have confidence in the interpretations of results on a widely used measure like the FCI and want to know if observed differences between females and males are artifacts of test bias or primarily due to factors such as background, attitude, or instruction. The findings from this study provide evidence that the FCI is not systematically biased in favor of males. However, three items were found to exhibit substantial DIF (two in

favor of males and one in favor of females) and warrant continued attention. If further investigation of possible sources of instrument bias and DIF analyses on other samples continue to provide evidence that the FCI is not systematically biased in favor of males, educators and researchers may continue using the FCI as a tool for quantifying gender differences, and as a valid source of evidence for assessing the effectiveness of interventions that promise to reduce the gender gap.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for Educational and Psychological Testing.* Washington, DC: American Educational Research Association.

Camilli, G. (2006). Test fairness. In R. L. Brennan (Ed.), *Educational Measurement,* (4th ed., pp. 221-256). Westport, CT: American Council on Education/Praeger.

Docktor, J., & Heller, K. (2008). Gender differences in both Force Concept Inventory and introductory physics performance. In C. Henderson, M. Sabbella, and L. Hsu (Eds), *2008 Physics Education Research Conference Proceedings*. Rochester, NY: American Institute of Physics.

Hake, R. (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.la.asu.edu/R&E/Research.html.

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

Holland, P. W., & Wainer, H. (Eds). (1993). *Differential Item Functioning*. Hillsdale, NJ: Lawrence Erlbaum.

Kahle, J. B., & Meece, J. (1994). Research on gender issues in the classroom. In D. L. Gabel (Ed.), *Handbook of Research on Science Teaching and Learning* (pp. 542-557). New York: Macmillan.

Kost, L. E., Pollock, S. J., & Finkelstein, N. D. (2009). Characterizing the gender gap in introductory physics. *Physical Review Special Topics – Physics Education Research, 5*, 010101, p. 1-14.

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.

Lorenzo, Crouch, & Mazur, (2006). Reducing the gender gap in the physics classroom. *American Journal of Physics, 74*, p. 188-122.

McCullough, L. (2004). Gender, context, and physics assessment. *Journal of International Women's Studies, 5*(4), 20-30.

McCullough, L. , & Meltzer, D. E. (2001). Differences in male/female response patterns on alternative-format versions of FCI items. In S. Franklin and K. Cummings (Eds), *2001 Physics Education Research Conference Proceedings*. Rochester, NY: American Institute of Physics.

Mullis, I. V. S., Martin, M. O., Fierros, E. G., Goldberg, A. L., & Stemler, S. E. (2000). *Gender Differences in Achievement; IEA's Third International Mathematics and Science Study (TIMSS)*. Chestnut Hill, MA: TIMSS International Study Center, Boston College.

Pollock, S. J., Finkelstein, N. D., & Kost, L. E. (2007). Reducing the gap in the physics classroom: How sufficient is interactive engagement? *Physical Review Special Topics – Physics Education Research, 3*, 010107, p. 1-4.

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.

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

Thornton, R. K., & Sokoloff, D. R. (1998). Assessing student learning of Newton's laws: The Force and Motion Conceptual Evaluation and the evaluation of active learning laboratory and lecture curricula. *American Journal of Physics, 66,* p. 338 - 352.

Wainer, H., Sireci, S. G., & Thissen, D. (1991). Differential testlet functioning: Definitions and detection. *Journal of Educational Measurement, 28*(3), 197-219.

Wang, W.-C., (2009). Assessment of Differential Item Functioning. In E. V. Smith, Jr. and G. E. Stone (Eds), *Criterion Referenced Testing: Practice Analysis to Score Reporting Using Rasch Measurement Models*. Maple Grove, MN: JAM Press.

Wilson, M. (2005). *Constructing measures: An item response modeling approach.* Mahwah, NJ: Lawrence Erlbaum Associates.

Wilson, M., & Iventosch, L. (1988). Using the partial credit model to investigate responses to structured subtests. *Applied Measurement in Education, 1*(4), 319 – 334.

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