Gender Bias In The Force Concept Inventory?

R.D. Dietz, R.H. Pearson, M.R. Semak and C.W. Willis

University of Northern Colorado, Greeley, CO 80639

Abstract. Could the well-established fact that males tend to score higher than females on the Force Concept Inventory (FCI) be due to gender bias in the questions? The eventual answer to the question hinges on the definition of bias. We assert that a question is biased only if a factor other than ability (in this case gender) affects the likelihood that a student will answer the question correctly. The statistical technique of differential item functioning allows us to control for ability in our analysis of student performance on each of the thirty FCI questions. This method uses the total score on the FCI as the measure of ability. We conclude that the evidence for gender bias in the FCI questions is marginal at best.

Keywords: Physics education research, gender, bias, student understanding, mechanics, differential item functioning
PACS: 01.40.Fk, 01.40.gf, 01.75.+m

INTRODUCTION

That males outperform females on the Force Concept Inventory (FCI) is as undeniable as is our perplexity as to why this should be the case [1]. One possibility is that the FCI is somehow infected with gender bias. Most of the thirty questions that comprise the FCI do not display any overt bias towards males, but some of the questions do deal with the motion of objects typically considered masculine such as hockey pucks and cannon balls. McCullough [2] has reported on the use of a revised FCI in which such masculine objects were replaced by more feminine ones. That modification did not result in any significant change in the average score attained by females on the test. Our investigation is intended to reveal gender bias of a more subtle nature.

Our own unpublished studies have shown that males outperform females on every one of the individual FCI questions (Figures 1 and 2). However, those analyses did not explicitly address the question of bias. Modern statistical approaches to check for item bias investigate whether examinees who have the same ability but are members of different subgroups have equal probability of answering a question correctly. This methodology is known as differential item functioning (DIF).

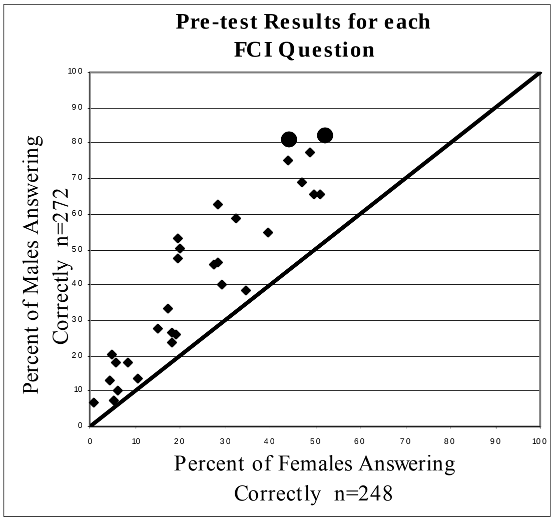

FIGURE 1. Comparison of male and female performance on the FCI given as a pre-test during the first week of the semester. Each point corresponds to one of the thirty FCI questions. Large circles indicate Questions 6 and 12 on the FCI which the statistics (discussed below) suggest are biased in favor of males.

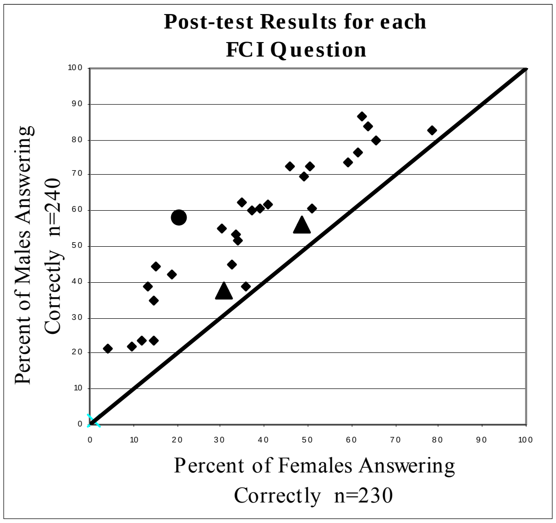

FIGURE 2. Comparison of male and female performance on the FCI given as a post-test during the last week of the semester. Each point corresponds to one of the thirty FCI questions. The large circle indicates Question 23 on the FCI which the statistics (discussed below) suggest is biased in favor of males. The triangles indicate Questions 4 and 9 which the statistics suggest are biased in favor of females.

METHODOLOGY

One method for studying DIF is through item response theory (IRT) which is a measurement framework that allows both examinee ability and item difficulty to be independently estimated from a set of test scores. In the IRT approach, the item characteristic curves for males and females would be compared for a given item [3]. Though theoretically appealing, one drawback to IRT-based methods is they require large sample sizes (perhaps 500 to 1,000 or more per group) for stable parameter estimates.

Another technique for studying DIF is the Mantel-Haenszel (M-H) method [4]. In this method, examinees are first stratified by some measure of ability, usually the total test score. Then for a given item the null hypothesis being tested is that there is no association between gender and score in any of the strata. This procedure does not require samples to be as large as the IRT-based methods. Hambleton & Rogers [5] demonstrated that the M-H and IRT-based DIF procedures tend to identify the same items as exhibiting DIF, provided that the DIF is uniform across ability.

TABLE 1. The grouping of students into strata for the pre- and post-tests – The most capable students are in stratum 5.

Pre-Test

Stratum	Male(%)	Female (%)	Total
1(low)	17 (17.7)	79 (82.3)	96
2	52 (38.2)	84 (61.8)	136
3	35 (46.1)	41 (54.0)	76
4	83 (71.6)	33 (28.5)	116
5(high)	85 (88.5)	11 (11.5)	96
Total	272	248	520

Post-Test

Stratum	Male(%)	Female (%)	Total
1(low)	29 (33.0)	59 (67.1)	88
2	27 (28.7)	67 (71.3)	94
3	40 (36.7)	69 (63.3)	109
4	61 (70.1)	26 (29.9)	87
5(high)	83 (90.2)	9 (9.8)	92
Total	240	230	470

In the present study, we used the M-H procedure to investigate DIF in the 30 items of the FCI. Students were administered the FCI before (as a pre-test) and after (as a post-test) completing a one-semester introductory physics course which emphasized mechanics. Pre- and post-test examinees were grouped into five strata based on their FCI total score using the cohort quintiles (see Table 1). Note that the percent of examinees in each stratum differs from 20% due to the discrete nature of FCI scores.

To calculate the M-H statistic, the odds for answering an item correctly (i.e. the probability of a correct answer divided by the probability an incorrect answer) are found for males and females by strata. Then a weighted sum of the odds ratios across the strata is computed. The resulting statistic can be thought of as an estimate of the overall ratio of the odds of correct response for females versus males after accounting for ability [6]. This statistic follows a chi-square distribution with one degree of freedom. A reduced significance level of 0.005 was used to control the increased chance of type I error associated with testing all 30 items for a given exam (pre or post). Analysis was performed using the FREQ procedure in SAS 9.2 after stratification by PROC RANK.

ANALYSIS

Pre-test and post-test items that exhibited significant DIF (at $\alpha = 0.005$) are presented in Table 2 with estimated odds ratios, 99.5% confidence limits in parentheses, and M-H chi-square statistics with

TABLE 2. Statistics concerning items exhibiting significant DIF on the pre- and post-tests

Item	OR (99.5% CI)	Chi-Sq	p-val	Favored
Pre-6	0.47 (0.25, 0.90)	10.60	0.0011	Males
Pre-12	0.37 (0.20, 0.70)	20.00	< 0.0001	Males
Post-4	2.07 (1.03, 4.15)	8.94	0.0028	Females
Post-9	2.26 (1.07, 4.78)	9.80	0.0017	Females
Post-23	0.43 (0.21, 0.86)	11.50	0.0007	Males

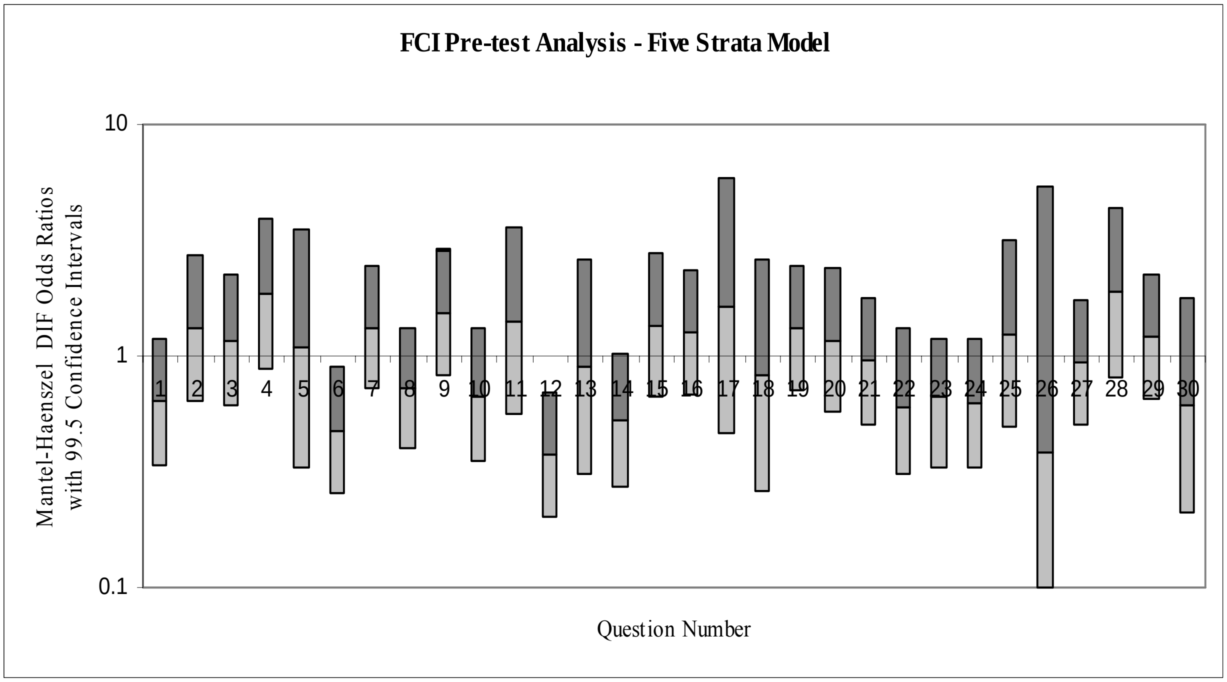

FIGURE 3. Mantel-Haenszel DIF Odds Ratio for each question when the FCI was administered as a pre-test. The length and location of each bar shows the extent of the 99.5 % confidence interval. Bias is suggested if the confidence interval does not include unity.

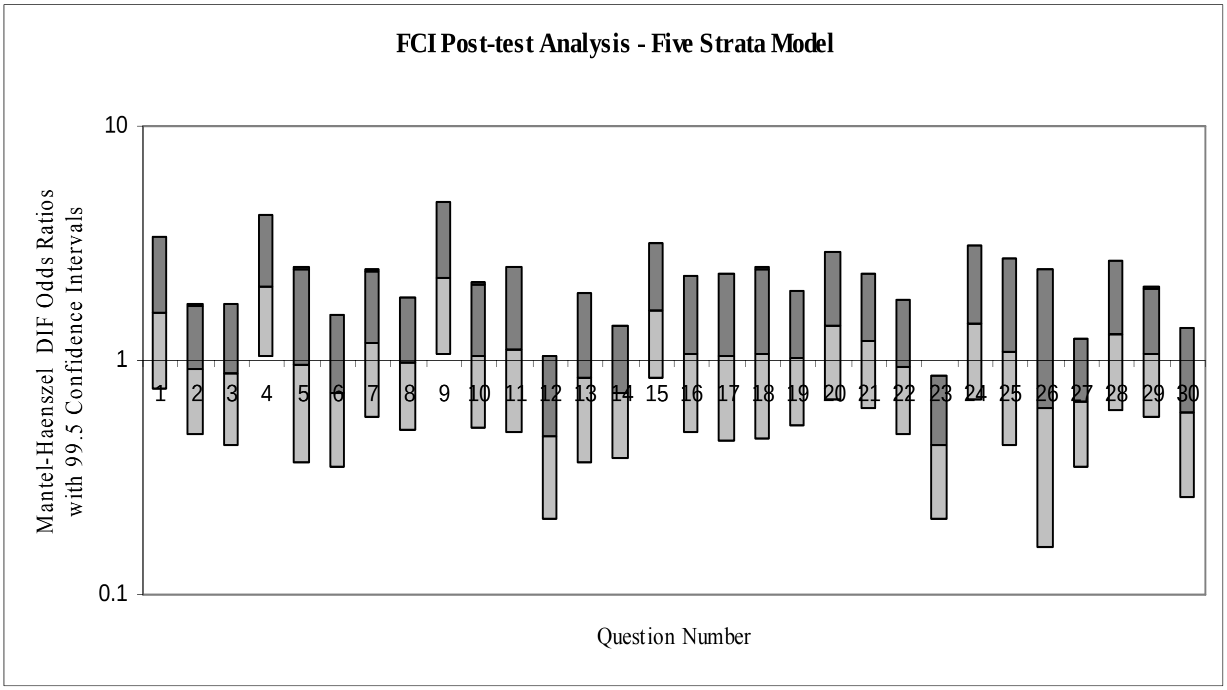

FIGURE 4. Mantel-Haenszel DIF Odds Ratio for each question when the FCI was administered as a post-test. The length and location of each bar shows the extent of the 99.5 % confidence interval. Bias is suggested if the confidence interval does not include unity.

corresponding p-values. Females are arbitrarily used as the reference group in this analysis, so odds ratios greater than one favor females, and those less than one favor males.

From Figures 3 and 4 above, it can be seen that, while most questions tend to favor one gender or the other, the 99.5% confidence level tends to overlap both genders. On the pre-test, only two questions fall totally within one gender. On the post-test, three different questions totally fall within one gender.

The results presented may seem paradoxical. How can it be that males do better than females on a question and, yet, the DIF results indicate that the question favors females? The following simple two-stratum example shows how this can be.

We divide a group of 100 females and 100 males into two equal strata based on their total test performance and consider their results on one test question. Table 3 shows the hypothetical breakdown of the number of males and females in each stratum answering the test item correctly (along with the corresponding percentages). Overall, males tend to do much better on the overall test, so the top stratum is predominantly male. For the high stratum, the odds that a female will answer the test item correctly are 9/1 whereas they are 7/3 for a male. The female-to-male ratio of these odds is 27/7 = 3.86. The same is true for the low stratum, thus, giving an average odds ratio of 3.86 which favors females. Yet, when the two strata are added, we see that males have given the correct answer more often than the females.

TABLE 3. Data for a simple two-stratum example – The third column gives the number of individuals in the stratum that answered the item correctly. The fourth column gives the percent of individuals that gave the correct answer.

Stratum	Population		#Correct		%Correct	
	M	F	M	F	M	F
Low	20	80	2	24	10	30
High	80	20	56	18	70	90
Total	100	100	58	42		

CONCLUSION

Our students' results on the FCI are similar to those found by a large number of other researchers in that males as a whole tend to outperform females (recall figures 1 and 2). This would tend to imply there is some kind of gender bias inherent in the test. To look more closely at this possible gender bias we performed a differential item functioning (DIF) analysis.

DIF analysis of our data draws attention to five questions on the FCI that appear to favor one gender over the other. We recognize the limitations posed by the size of our data set, which took four years to collect. Our results would be on firmer statistical grounds if we had larger classes.

Have we gathered enough evidence to suggest that some questions be removed from the FCI because of bias? No. Statistical arguments alone do not suffice. They only cast suspicion on certain questions. There must also be plausible evidence based on the wording and/or context of the question to conclude that bias exists. The one possible exception to this verdict is the question that shows the strongest DIF, Question 12, which favors males and involves a cannon. A plausible but not exactly probable case can be made in this lone instance.

REFERENCES

1. J. Docktor and K. Heller, "Gender Differences in Both Force Concept Inventory and Introductory Physics Performance," *2008 Physics Education Research Conference,* AIP Conf. Proceedings 1064, AIP, Melville, NY, 2008, pp. 15-18.
2. L. McCullough, "Gender, Context, and Physics Assessment" *J. Int. Women's Studies* **5**, 20 (2004).
3. R. K. Hambleton, H. Swaminathan and H. J. Rogers, *Fundamentals of Item Response Theory*, Newbury Park: Sage Publications, 1991.
4. G. Camilli and L. A. Shepard, *Methods for Identifying Biased Test Items: Volume 4*, Thousand Oaks, Sage Publications, Inc., 1994.
5. R. K. Hambleton and H. J. Rogers, "Detecting potentially biased test items: Comparison of IRT area and Mantel-Haenszel methods," *Applied Measurements in Education* **2**, 313 (1989).
6. L. Roussos, D. Schnipke and P. Pashley, "A Generalized Formula for the Mantel-Haenszel Differential Item Functioning Parameter," *J. Educational and Behavioral Statistics* **3**, 24 (1999).