UNIVERSITY *of* NORTHERN COLORADO

Gender Bias and the Force Concept Inventory

by

Richard D. Dietz, Robert H. Pearson, Matthew R. Semak, Courtney W. Willis
University of Northern Colorado, Greeley, CO 80639

Abstract: Is the well-established fact that males tend to outperform females on the Force Concept Inventory (FCI) evidence of gender bias? A question is biased only if factors other than ability determine a student's performance on that question. Using the total score on the FCI as a proxy for ability, we describe and employ a variety of statistical techniques in an effort to identify which questions on the FCI may be accused of gender bias.

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. One possibility is that the FCI is somehow infected with gender bias. A careful inspection of the thirty questions that comprise the FCI does not disclose any overt bias towards males, but perhaps statistical analysis of test results can identify problematic questions.

Our own studies (see right) show that males outperform females on ALL of the individual FCI questions. However, this analysis is simplistic in that it does not take individual ability into account. A more valid way to check for bias asks whether males and females of equal ability (as determined somehow), will have equal probabilities, or odds, of answering a question correctly.

An area of statistics called differential item functioning (DIF) will do just this. With this method, only groups of subjects with comparable ability, termed strata, are compared.

For our analysis, the strata are determined by overall test scores. Given our sample size, we found it optimal to use lumped scores and divide the 31 possible outcomes into five strata.

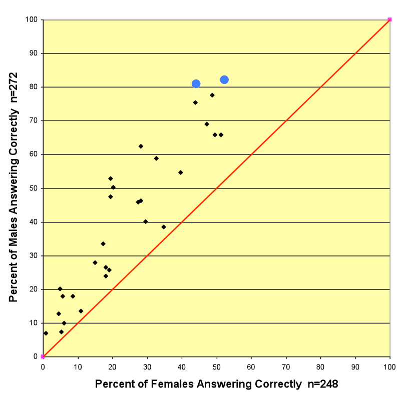
Analysis

In our DIF analysis, the odds for answering each test question correctly were calculated for females and males by strata. To show any differences in these odds, a stratum dependent odds ratio of females-to-males was then constructed. Finally, a weighted (by population per stratum) average was performed over the strata giving what is known as the Mantel-Haenszel odds ratio for each question.

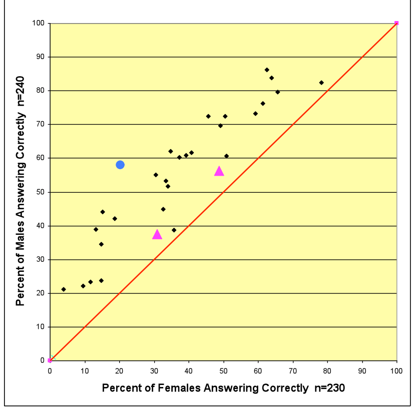
With this female-to-male ratio, values greater than 1 seem to favor females while values less than 1 tend to favor males. To determine whether that value is significant we chose a high 99.5% confidence ratio.

From the graphs to the right 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 fell totally within one gender. On the post-test three different questions totally fall within one gender.

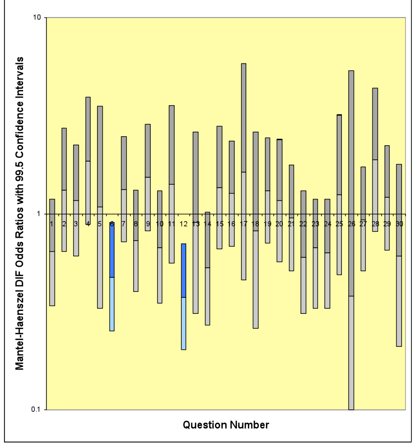
Pre-test Results for each FCI Question



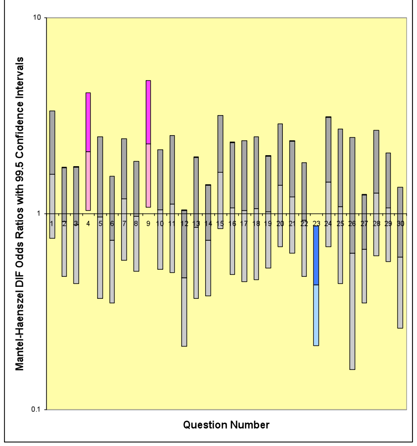
Post-test Results for each FCI Question



FCI Pre-test Analysis - Five Strata Model



FCI Post-test Analysis - Five Strata Model



References

1. L. K. Muthén and B. O. Muthén, *Mplus User's Guide 5th Ed.* Los Angeles: Muthén & Muthén (1998-2009).
2. G. Camilli, L. Shepard, *Methods for Identifying Biased Test Items* Thousand Oaks, California, SAGE Publications (1994).

Explanation

The results presented to the left 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-strata 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.

Overall, males tend to do much better on the test, so the top stratum is predominantly male. In each stratum the females outperform the males by 20 percentage points, yielding an odds ratio of 3.85 which favor females. Yet, when the two strata are added we see that the males have given the correct answer more often than the females.

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

Conclusions

Our students' results on the FCI are similar to a large number of other researchers in that males as a whole tend to out perform females (see top two graphs.) 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 used differential item functioning (DIF) .

The average DIF odds ratio for both pre and post tests was less than 1.1 ,but the standard deviation for both tests was above 0.4 which indicates no statistical bias. 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 more data.

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.