What's So Special about Question 23?

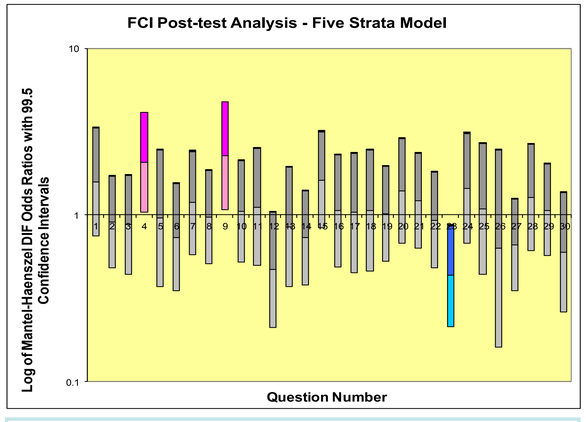
Richard D. Dietz, Wendy K. Adams, Matthew R. Semak, Courtney W. Willis
University of Northern Colorado, Greeley, CO 80639

Abstract: Recent gender studies employing Differential Item Functioning (DIF) have shown that Question 23 on the Force Concept Inventory exhibits significant DIF in favor of males. Question 23 is the third in a quartet of questions that examine the motion of a rocket in outer space. We present further analysis of the responses to these four questions in an effort to determine why this particular question exhibits gender bias.

Introduction

Most researchers who have used the Force Concept Inventory (FCI) as a metric for gauging student understanding of introductory mechanics have observed that, on average, males outperform females. This is a robust result in spite of the fact that other metrics (such as grade in the course) may show that males and females perform at the same level. Such findings may lead one to ask whether the FCI harbors gender bias. At least two studies utilizing the statistical technique of differential item functioning (DIF) have examined each question on the FCI for evidence of gender bias. DIF divides test-takers into two groups (male and female in this case) and within each group divides the students into several strata ranked by total score on the FCI as a proxy for ability. Gender bias in a particular question would exist if males of a given ability level outperformed females at the same ability level by a statistically significant amount.

In this presentation we reference two unpublished studies that have used DIF to uncover gender bias in FCI questions and that have concluded that question 23 is biased in favor of males. The study by Osborn Popp et al. at Arizona State Univ. examined the responses of 4775 high school physics students on the FCI given as a post-test. Our group at the University of Northern Colorado (UNC) has collected post-test results from 396 students over a period of four years. In both studies the numbers of males and females who participated are roughly equal.



DIF Analysis

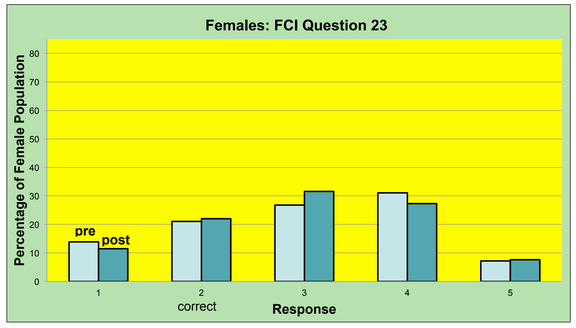
In our DIF analysis, the odds for answering each test question correctly were calculated for females and males by strata. To show any differences in these odds, a stratum dependent odds ratio of females-to-males was then constructed. Finally, a weighted (by population per stratum) average was performed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

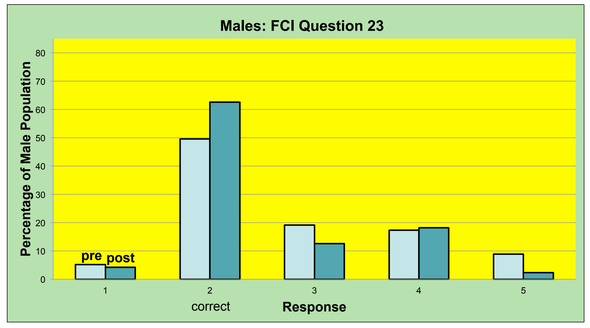
With this female-to-male ratio, values greater than 1 seem to favor females while values less than 1 tend to favor males. To determine whether that value is significant we chose a high 99.5% confidence ratio.

From the graph 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 post-test, three different questions totally fall within one gender. In particular, Q23 tends to favor males.

Gathering Insight

It is interesting to explore female and male choices of the possible responses to each FCI question. Using our data, gender specific percentages for the five responses of each question on the FCI, pre and post, were recorded. This information is presented visually using bar-graphs which show each question's response profile, pre and post, for each gender. In particular, the results for Q23 are shown in the plots below.





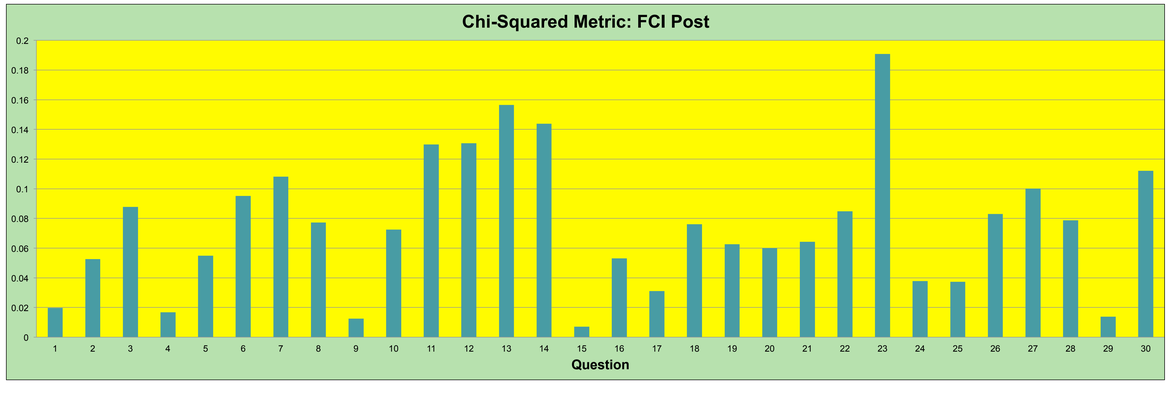
The plots above show the gender specific response profiles for Q23. This question shows the most pronounced profile difference between the genders.

Quantifying Profile Differences

In comparing these profiles by gender, those of Q23 suggest the starkest difference. This is confirmed, quantitatively, for the post-test scenario using the chi-squared metric,

$$D(P_F,P_M)=\frac{1}{200}\sum_{i=1}^{5}\frac{(P_F(i)-P_M(i))^2}{P_F(i)+P_M(i)}$$

P_F and P_M are the percentage of females and males, respectively, that selected a particular response, and the index, i, ranges over the five possible responses. Of course, this is done separately for pre and post data. As scaled, $0 \le D \le 1$, where larger values tell of a greater degree of mismatch between profiles. Our post-test profile comparisons are shown below.



The graph above shows the result of quantifying the difference between FCI (post) question profiles, by gender, using the chi-squared metric. Q23 shows the greatest mismatch between profile pairs.

In general, female and male question profile pairs often similar with males selecting the correct response more often.

This is certainly not true for Q23. The female profile for Q23 shows a quasi-uniform distribution of percentages with a (not very pronounced) maximum on the incorrect response. Moreover, there is virtually no evolution from the pre to post-test results. The male profile for Q23 has a rather strong peak (pre and post-test) on the correct response.

- Our general classroom experience does not track with the overall FCI gender performance difference much less such results as seen on Q23.
- We feel our quantitative work has served us well in indicating that FCI questions such as Q23 require further investigation as to their role in properly gauging student performance.
- To pursue an answer as to why students react to such questions as they do, we turn to conducting student interviews the preliminary results of which are detailed in the third column.

Student Interviews

We are embarking on a series of student interviews to better understand our results.

General Results:

❖Students are picking the incorrect *and* correct answers for a wide range of reasons.

❖Scores have not been consistent with the depth of student understanding.

Four student interviews have been conducted so far (2 male, 2 female). All completed two semesters of introductory physics (2010-2011).

Results *(Too soon to generalize):*

- None of the students guessed
  - All chose an option with physical reasoning behind their answers
- Scores on 18 questions ranged from 28% - 56%
- Lowest score
  - Female
  - Used 100% real world reasoning
  - Consistent and most *logical* reasoning of all 4 students
  - Difficulty with vocabulary
  - Difficulty with "common physics scenarios"
- Highest score
  - Male
  - Used 100% classroom reasoning – not experience
  - Superficial logic – difficulty applying concepts/logic
  - Excellent understanding of velocity, acceleration (1-D) and Newton's 1st Law
  - At ease with "common physics scenarios"

Specific results for Questions (21-24)

- Scores: Q21: 2, Q22: 1, Q23:2, Q24: 3
- Many different reasons were given for incorrect answers

Two exceptions:

- Q23: Two students did not picture the correct direction approaching point c.
- Q22: Two students visualized thrust turned on a bit and then turned off. One other did this initially but corrected herself. The fourth sees all motion the same, i.e., velocity = acceleration.

- Two students were unsure about "space" and what rules apply there.

Interview Protocol

- 18 FCI questions
- ~1 hour
- Think aloud in two phases
  - Introduction: Ask if they remember taking this test in lab. Tell them we don't understand the results, because they don't seem to fit with what we see in class.
  - Phase 1: Ask student to answer each question and think out loud.
  - Phase 2: Tell student that we're going back over these questions, and I'll tell you the Physicist's answer. Have student explain why s/he thinks that the physicist might choose that answer, and if it seems reasonable.
  - Close: Go back and tell students where they were off

Interesting Note: Women seem to be less willing than men to volunteer for an interview that involves physics content.