Gender Bias in the FCI - Comparison Across Studies

Matthew R Zachreson

BYU IDAHO Department of Physics

Introduction

The FCI has been used to assess student learning for 30 years. During that time, many have documented a “gender gap” - a measurable difference in performance between male and female students (see Madsen et al [1] for an overview).

In an attempt to understand why, several researchers have turned to item responses theory (IRT) to analyze the differences in how males and females perform on the questions, including Osborn Popp et al [2] and Traxler et al [3].

This paper replicates their studies using data collected at BYU-Idaho in order to compare all three results side by side.

The Data

- Osborn Popp et al [2] analyzed a sample of 4,775 tests collected from high schools as participating in workshops sponsored by the national science foundation. IRT analysis was done using a Rasch model (1 parameter fit)
- Traxler et al [3] analyzed a sample of 9,228 tests collected from large, southern land-grant university enrolling approximately 25,000 students. IRT analysis was done using a two parameter fit.
- This paper analyzes 10,504 test results collected at BYU-Idaho, a private, non-profit, primarily undergraduate institution in south-east Idaho enrolling approximately 25,000 students. IRT analysis was done using a two parameter fit.
- For all three samples, the number of students identifying as nonbinary or something other than male/female was too small for statistical significance and they had to be excluded from the study.

	Male students		Female students		
	N	$(M \pm SD)\%$	N	$(M \pm SD)\%$	d
Osborn Popp					
Pretest	N/A				
Post-test	2427	59 ± 22	2348	$45 \pm 20.$	0.65
Traxler					
Pretest	3482	43 ± 18	1027	32 ± 14	0.69
Post-test	3628	73 ± 17	1088	65 ± 18	0.46
This Paper					
Pretest	4401	45 ± 20	1170	30 ± 16	0.83
Post-test	3532	68 ± 22	966	53 ± 23	0.56

References

[1] A. Madsen, S. B. McKagan, and E. Sayre, Gender gap on concept inventories in physics: What is consistent, what is inconsistent, and what factors influence the gap?, Phys. Rev. ST Phys. Educ. Res. 9, 020121 (2013).

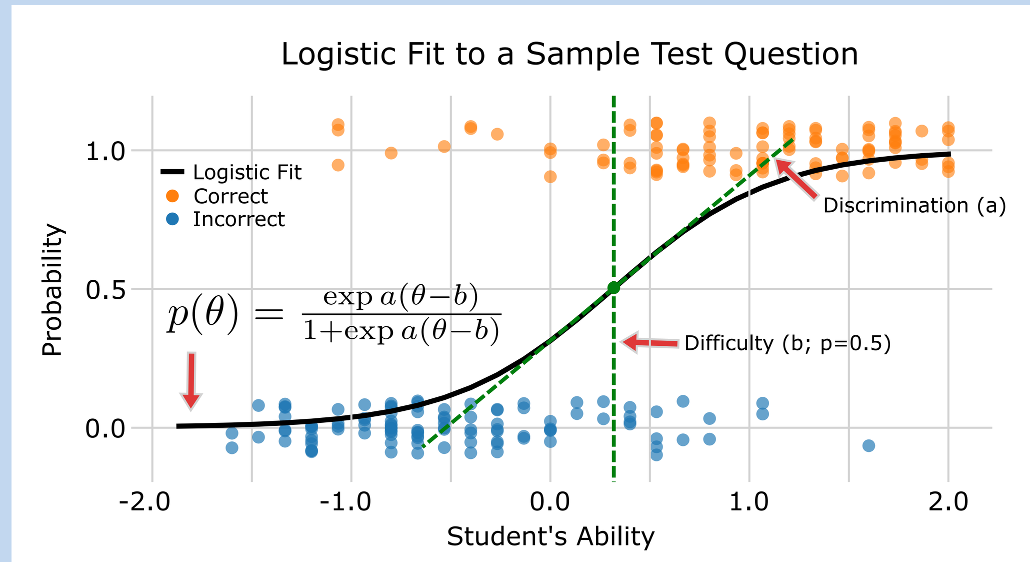
[2] S. Osborn Popp, D. Meltzer, and M. C. Megowan-Romanowicz, Is the Force Concept Inventory biased? Investigating differential item functioning on a test of conceptual learning in physics, in 2011 American Educational Research Association Conference (American Education Research Association, Washington, DC, 2011).

[3] A. Traxler, R. Henderson, J. Stewart, G. Stewart, A. Papak, and R. Lindell, Gender fairness within the force concept inventory, Phys. Rev. Phys. Educ. Res. 14, 010103 (2018).

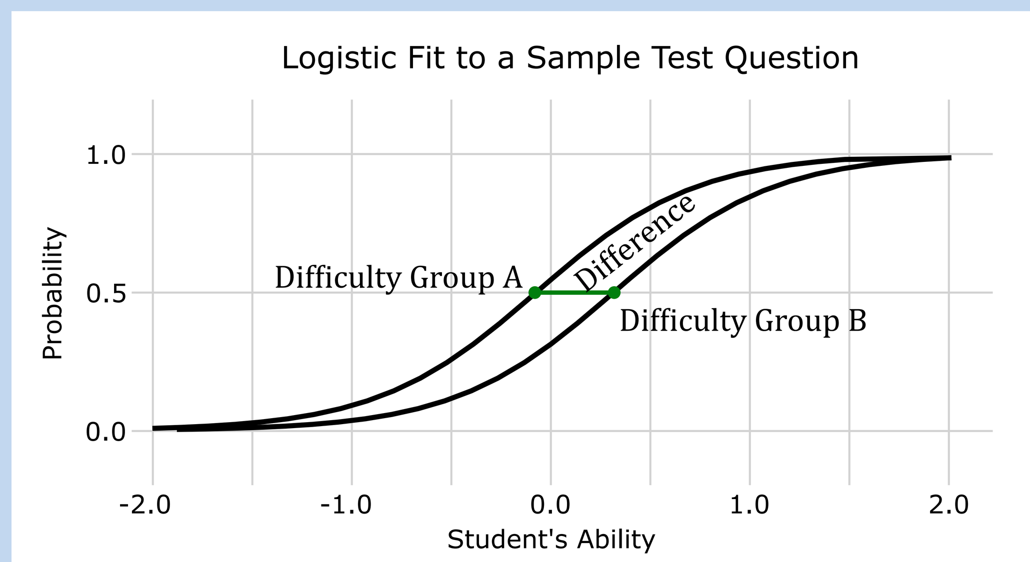
[4] Philip Eaton and Shannon D. Willoughby, Confirmatory factor analysis applied to the force concept inventory, Phys. Rev. Phys. Educ. Res. 14, 010124 (2018).

Item Response Theory

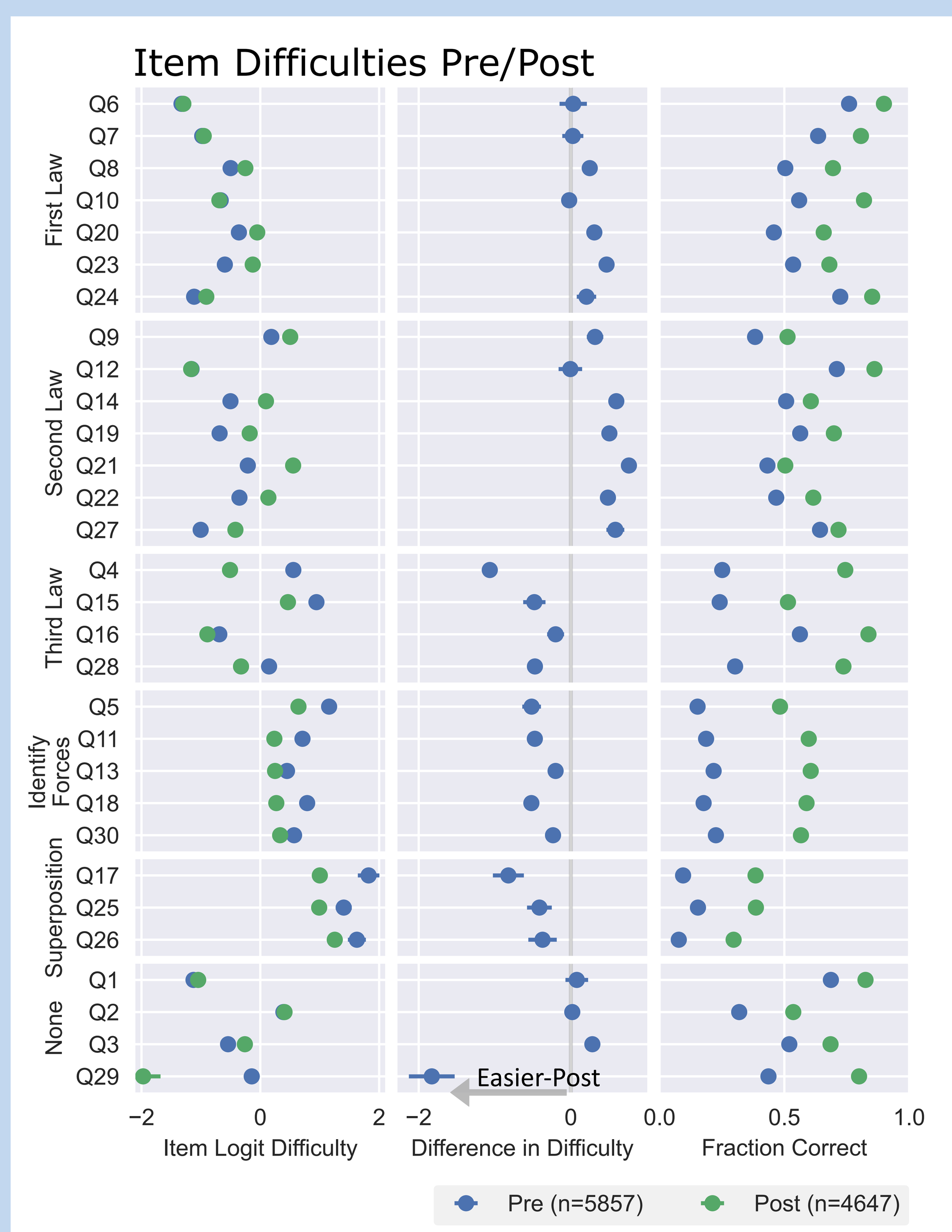
Uses logistic regression to find the probability that a student will answer a question correctly, given the student’s ability level:



Differential Item Functioning (DIF) compares those fits to each other, answering the question “Will two students of the same ability but from different groups answer the question the same way?”

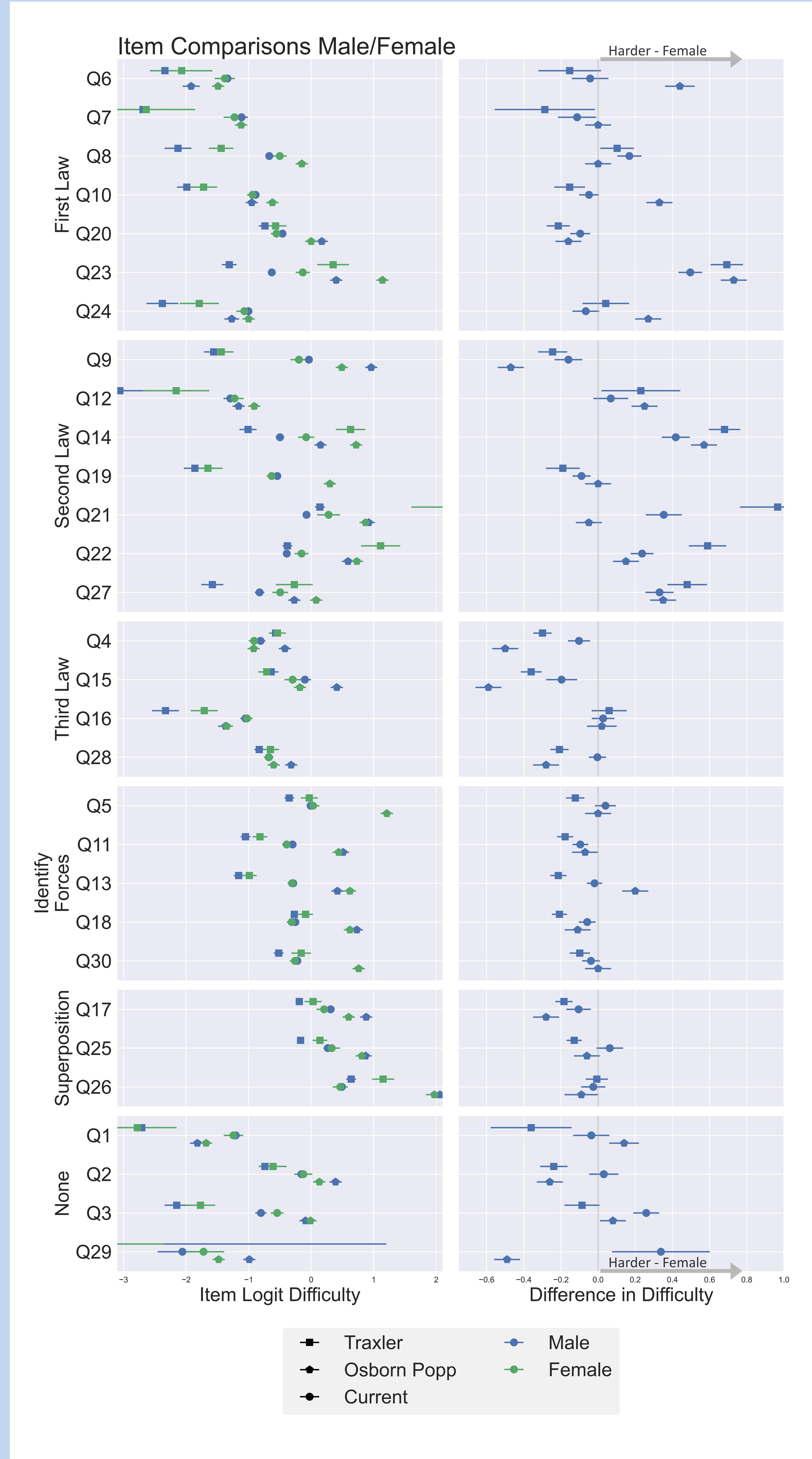


Example comparing pre-test scores to post-test scores. They are grouped by the factors found by Eaton and Willoughby [4] in their EW5 model.



Notice that Third Law problems are relatively easier on the post test than on the pre-test. That means a student scoring 15/30 on the post test is more likely to answer the third law questions correctly than a student scoring 15/30 on the prestest.

Results



To do DIF, the test scores need to be anchored to allow the comparison between groups. Osborn Popp [2] anchored by assuming that the overall test score was indicative of student ability. Traxler [3] assumed that both groups had equal ability and any differences were due to the test. This paper did both and used the comparison between both results to scale Traxler’s results to match the same scale as Osborn Popp, allowing for easy visual inspection of any trends.