The Impacts of Instructor and Student Gender on Student Performance in Introductory Modeling Instruction Courses

Daryl McPadden[α], Eric Brewe[αβγ]

Florida International University, 11200 SW 8th St., Miami, FL 33199
[α] Department of Physics, [β] Department of Teaching and Learning,
[γ] STEM Transformation Institute

Abstract. This study considers the impact of instructor on the gender gap in students' scores on the Force Concept Inventory (FCI) in Modeling Instruction (MI) courses at Florida International University (FIU). Earlier work has shown that MI had increased FCI scores overall when compared to traditional lecture courses; however, the gap between male and female students' scores in the MI courses increased over the course of the semester. Student data were collected from 559 students at FIU, over 18 semesters, with 10 different instructors. General linear regression was used to determine the significance of the student gender and instructor factors in predicting a student's FCI score post-instruction and the fraction of variance explained by these factors. Effect sizes were then calculated from the difference in female students' gains from male students' gains and compared between instructors. Analysis showed an instructor-independent, medium effect favoring male students' scores on the FCI.

INTRODUCTION

It has long been a standard in physics that male students outperform female students, be it on classroom tests, course grades, or standardized concept inventories [1, 2]. This difference in scores is often referred to as the "gender gap". Recently, the gender gap on physics concept inventories was shown to be present across multiple different inventories and institutions, including both lecture-style and reformed-style teaching methods [2]

In particular, Brewe et. al found "Modeling Instruction (MI) widens the understanding gap on FCI scores that exists between men and women prior to instruction" at Florida International University (FIU)–a large, Hispanic-serving institution [3]. MI is a reformed curriculum and pedagogy where lecture, lab, and recitation are combined in a studio format, which encourages students' participation and agency through group activities and guided classroom discussions [4]. Students are responsible for developing, testing, evaluating, and revising models, which are defined as a coordinated set of representations that can be used to explain, predict and describe a particular class of phenomena [5]. In this classroom format, students' FCI scores, for both male and female students, were shown to be higher when compared to the equivalent lecture-based course. Despite the careful design and planning that went into the MI mechanics course, it is still plagued with the growth of the gender gap internal to the course, providing larger FCI gains for male students than for female students [3].

The purpose of this study is to explore the difference in FCI scores between male and female students in the introductory MI courses at FIU, using the existing FCI data archive at FIU, and to propose plausible explanations for the results found.

METHODS

FCI data were collected through 18 semesters of introductory mechanics MI courses, spanning from Fall 2004 to Fall 2013 [6]. During this time, 10 different instructors taught the course, bringing a variety of backgrounds and experiences to the classroom. Pre- and post- FCI scores were matched for 559 students, 257 of which were female and 302 of which were male. The number of matched scores for each instructor varied from 17 to 153, depending on the number of semesters the instructor taught and the number students in each section.

Histograms of students pre- and post-FCI scores, shown in Figure 1, reveal notably different distributions in post-scores when disaggregated by student gender. The pre-score distributions for both male and female students are similar and skewed toward the low end of the range; however, the post-score distribution for male students shows both an overall shift toward higher scores and the distribution itself is now skewed toward higher scores, while the post-score distribution for female students only shows an overall shift toward higher scores, maintaining the right skewed distribution seen in the fe-

male pre-scores. Similar distributions and shifts are seen when the data are separated by instructor.

General Linear Regression

In order to further investigate the difference in scores seen in Figure 1, general linear regression was used on the raw FCI scores (out of 30 points) to determine how much of the variance in students' post-FCI scores could be determined by the students' pre-FCI scores, genders, and instructors. Note that all instructors in the model were compared to Instructor A because Instructor A had both the most experience teaching MI and the greatest number of students overall. This makes Instructor A the natural baseline for comparison. A threshold of $\alpha = 0.05$ was used to determine significant results.

Effect Size

To quantify the difference in scores, the raw percentage gain ($\%Gain = (Post\% - Pre\%)$) was calculated. (Raw gain is used over normalized gain because of the non-linear nature of normalized gains.) Cohen's d was then calculated from the students' gains using

$$d = \frac{\mu_f - \mu_m}{\sigma_{pooled}} \quad (1)$$

where μ_f is the average gain for female students, μ_m is the average gain for male students, and σ_{pooled} is the pooled standard deviation given by

$$\sigma_{pooled} = \sqrt{\frac{(n_f - 1)s_f^2 + (n_m - 1)s_m^2}{n_f + n_m - 2}} \quad (2)$$

Here, n_f & n_m are the number of female and male students, and s_f & s_m are the standard deviations of the female and male gains, respectively [7, 8]. Cohen's d was calculated for the total data set as well as for each instructor. The 95% confidence interval (CI) was calculated from

$$CI = d \pm 1.96 * s_d \quad (3)$$

where d is from equation (1) and s_d is given by

$$s_d = \sqrt{\frac{n_f + n_m}{n_f * n_m} + \frac{d^2}{2(n_f + n_m)}} \quad (4)$$

where n_f and n_m are the same as from equation (2) and d is again from equation (1) [8]. Thus, Cohen's d serves as a measure of the magnitude of the gender gap in students' gains of FCI scores. As an effect size, Cohen's d allows for the comparison of the gender gap across instructors with varying numbers of students.

TABLE 1. Summary of general linear regression for post-FCI score as predicted by pre-FCI score, student gender, and instructor. All instructors listed are compared to Instructor A. The bold face p-values are those that were considered significant at $\alpha = 0.05$.

| Factor	Estimate	p-value
Pre-FCI Score	0.728	$\approx$ **0**
Male	2.504	$\approx$ **0**
Instructor B	-1.760	**0.001**
Instructor C	-2.498	**0.008**
Instructor D	-1.940	0.067
Instructor E	-3.062	$\approx$ **0**
Instructor F	-3.304	**0.004**
Instructor G	-1.565	**0.030**
Instructor H	0.649	0.315
Instructor I	-0.254	0.791
Instructor J	-5.400	$\approx$ **0**

ANALYSIS AND RESULTS

General Linear Regression

The results of the general linear regression are shown in Table 1. The independent variables (left column) are the contributing factors considered in this analysis, the middle column is the estimated beta values for each contributing factor, and the p-values (right column) are a measure of the significance for the variables. A p-value of less than 0.05 was considered significant.

The adjusted R^2 value for the model is 0.517, which means that the students' pre-FCI scores, genders, and instructors account for 51.7% of the variance in the students' post-FCI scores. When broken down for each factor, pre-FCI score accounts for 43.4% of the variance, student gender accounts for 3.1% of the variance, and instructor accounts for 5.2% of the variance.

As Table 1 indicates, both the student's gender and instructor are significant predictors for the student's post-score on the FCI, which leads to two different conclusions. First, MI courses are affecting female students differently from their male counterparts, in such a way that male students have a better conceptual understanding of the material as measured by the FCI after the course. Second, students' conceptual understanding, as measured by the FCI, is dependent on the instructor of the course, meaning student learning is influenced by more than just the curriculum. Since there were no significant interactions between student gender and instructor, this implies that the gender gap in post-scores, when controlled for pre-scores, is consistent across all instructors. There were no significant interactions between pre-scores and gender or pre-score and instructor, which means that all the students are starting the semester on fairly equal conceptual understanding, again as measured by the FCI.
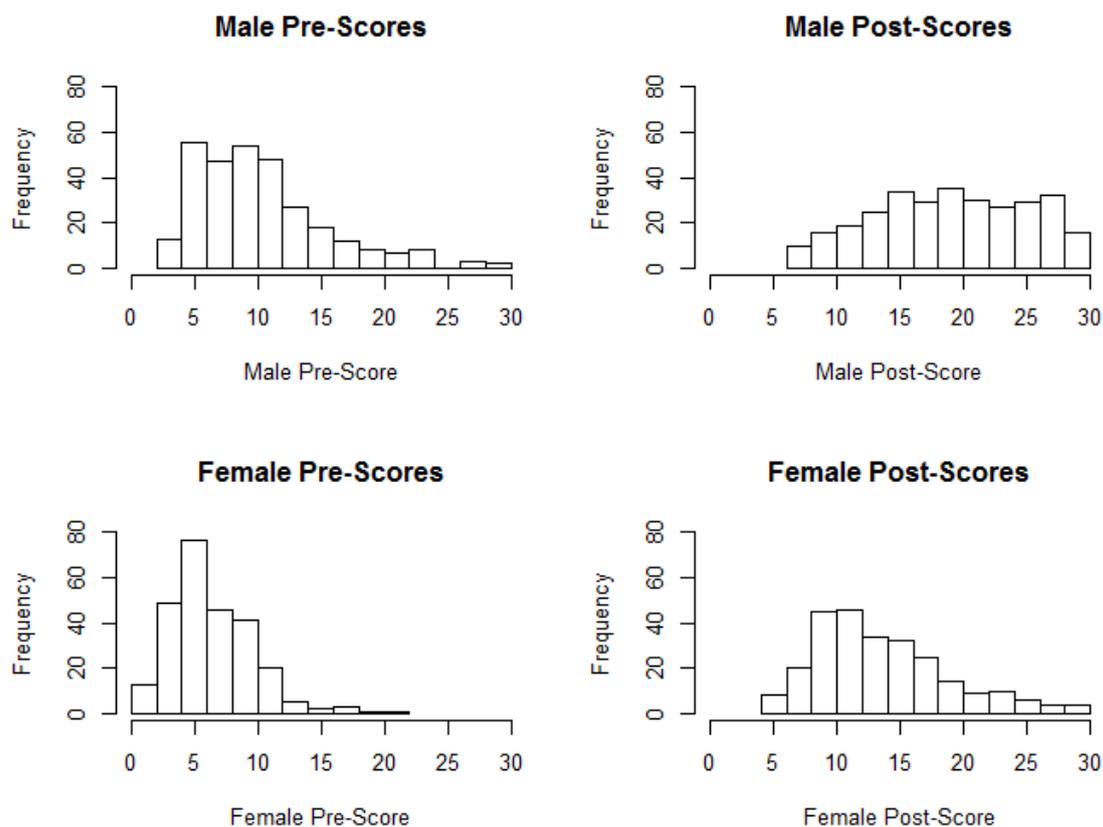
FIGURE 1. Histograms showing pre- and post-FCI scores for MI students, disaggregated by student gender.

Effect Size

The results of the effect size analysis are shown in Table 2. For each instructor and for the total data set, Cohen's *d* is calculated as a measure of the gender gap (second column from the right), the 95% CI on *d* is given (second column from the left), and the number of students for each calculation is shown (left column). A negative *d*-value means that male students have larger conceptual gains than the female students (as measured by the FCI), a *d*-value equal to zero means that the gains for male and female students are equal, and a positive *d*-value means that female students have larger gains than the males. The bold face data sets in Table 2 indicate those that have CIs which do not cross zero.

As is displayed by the histograms in Figure 1, we see a medium, negative effect of -0.311 for the total data set, which means that the average male student is outscoring approximately 67% of the female students. This confirms and quantifies the gender gap present in the MI course FCI gains. When broken down by instructor, only two out of the 10 instructors have CIs which do not cross zero; however, this is largely due to small sample sizes.

TABLE 2. Summary of effect size analysis as described by equations (1)-(4). The bold face data sets are those whose confidence intervals do not cross zero.

Data Set	d	CI (95%)	N
Instructor A	-0.399	± 0.320	153
Instructor B	-0.450	± 0.362	122
Instructor C	-0.023	± 0.804	26
Instructor D	-0.634	± 0.911	20
Instructor E	-0.044	± 0.530	59
Instructor F	-0.566	± 0.979	17
Instructor G	-0.320	± 0.555	51
Instructor H	-0.152	± 0.476	68
Instructor I	-0.208	± 0.783	25
Instructor J	-0.110	± 0.944	18
Total	-0.311	± 0.167	559

The other instructors all show negative effects, but the CIs are too large to draw supported conclusions. Since the CIs for all the data sets overlap, we see no instructor dependence in these gains, which follows the results found in the linear regression analysis (note that this could be due in part to small sample sizes for some

instructors).

From the effect size analysis, we conclude that the MI mechanics course produces larger gains on the FCI for male students than for female students, which is a medium-size effect that does not appear to be instructor dependent.

CONCLUSIONS AND DISCUSSION

From this analysis, we draw two main conclusions. First, MI courses affect male students differently than their female counterparts, in such a way that male students have larger gains on the FCI when compared to female students. This is evidenced by the distinct distributions in pre- and post-FCI scores (Figure 1), the significance of gender as a predictor of post-FCI score in the linear regression (Table 1), and the medium-size, negative effect calculated from FCI gains (Table 2). Second, while the raw post-FCI scores *are* instructor dependent, we see that the gender gap in FCI scores *is not*. This is demonstrated by the significance of instructor as a predictor of post-FCI scores, combined with the lack of significant interaction predictors between instructors and student gender in the linear model. In addition, the CIs of Cohen's *d* for all instructors overlap (Table 2), which means that in terms of the gender gap on FCI gains, the instructors cannot be distinguished from one another. Through these results, we have highlighted the issue of a growing gender gap on FCI scores in MI courses and quantified the effect seen in existing data.

These results lead to the question, "why do we see a growing gender gap in MI courses?" We propose four possible contributing factors to the gender gap in MI mechanics courses. First, McCullough found evidence of a gender bias within the FCI test itself [9]. Perhaps, after working within the physics community for a semester (as a strongly male-biased field), female students are more susceptible to the gender bias of the FCI, leading to a growth in the gap between male and female students' scores. Furthermore, stereotype threat often causes female students to perform worse on tests than their male counterparts [1]. Second, Kost, Pollack and Finkelstein found that background factors such as math achievement and high school physics preparation were able to explain the persistence of the gender gap [1]. Since students' pre-scores account for 43.4% of the variance in post-scores, we hypothesize that students' backgrounds could largely contribute to the gender gap at FIU (students' background factors were not available for this analysis). Third, the MI curriculum was primarily male-developed and primarily male-taught [4, 10, 11], which could have led to an implicit gender bias within the materials and structure of the course itself. In addition, Carlone found that high-achieving female students resisted reformed classes because it was so different from their previous learning environment, in which they thrived [12]. If this were the case, those high-achieving female students would have lower scores than expected, skewing the data toward lower scores (as is seen in Figure 1). Fourth, there are many instructor variables, such as years of teaching experience, teaching style, or instructor gender, which influence how students perceive the instructor and how much they are willing to learn [13]. However, given that the gender gap was not found to be instructor dependent, it is unlikely that instructor variables are the primary influential factors.

It is impossible to address this new driving question and its possible contributing factors with the current data set of FCI records. Interviews with students, learning assistants, instructors, and the curriculum developers would enrich the data set and allow a more deliberate focus on why we are seeing the gender gap growth that we do. In particular, the interviews should focus on students' backgrounds as pre-score was the largest explanatory factor. This is the future direction of the study.

ACKNOWLEDGEMENTS

REFERENCES

1. L. E. Kost, S. J. Pollock, and N. D. Finkelstein, *Phys. Rev. ST Phys. Educ. Res.* **5**, 010101 (2009).
2. A. Madsen, S. B. McKagan, and E. C. Sayre, *Phys. Rev. ST Phys. Educ. Res.* **9**, 020121 (2013).
3. E. Brewe, V. Sawtelle, L. H. Kramer, G. E. O'Brien, I. Rodriguez, and P. Pamelá, *Phys. Rev. ST Phys. Educ. Res.* **6**, 010106 (2010).
4. E. Brewe, *Am. J. of Phys.* **76** (2008).
5. S. Manthey, and E. Brewe, *Cell Bio. Ed. Life Sci. Ed.* **12**, 206–214 (2013).
6. D. Hestenes, M. Wells, and G. Swackhamer, *The physics teacher* **30**, 141–158 (1992).
7. I. Rodriguez, E. Brewe, V. Sawtelle, and L. H. Kramer, *Phys. Rev. ST Phys. Educ. Res.* **8**, 020103 (2012).
8. R. Coe, "It's the effect size,stupid: what effect size is and why it is important," 2002.
9. L. McCullough, *Journal of International Women's Studies* **5**, 20–30 (2004).
10. D. Hestenes, *Am. J. Phys.* **55**, 440–454 (1987).
11. I. A. Halloun, *Modeling theory in science education*, Springer, P.O. Box 17, 3300 AA Dordrecht, The Netherlands, 2006, ISBN 10140205114.
12. H. B. Carlone, *Journal of Research in Science Teaching* **41**, 392–414 (2004).
13. S. K. Bennett, *J. Ed. Psych.* **74**, 170–179 (1982).