Exploring the Gender Gap in One Department's Algebra-based Physics Course

Twanelle Walker Majors, Paula V. Engelhardt, Steve Robinson
Tennessee Technological University

Introduction

This preliminary work was conducted to determine if pedagogy (Traditional versus LEAP) and gender accounted for differences in performance on the Force Concept Inventory (FCI) and the Gender Force Concept Inventory (GFCI) for students enrolled in the 1st semester algebra-based physics course at Tennessee Tech University.

There are two versions of this course. The Traditional version meets three times a week for 55 minutes per class for lecture with a separate three-hour laboratory. The LEAP version meets for 2 hours three times a week for 120 minutes per class. LEAP sections use a guided-inquiry curriculum developed and supported through two NSF grants.

This portion of the work addressed the following questions:

1. Is there a difference in the performance of males and females on the Force Concept Inventory?
2. Is there a difference in the performance of males and females on the Gender Force Concept Inventory?
3. Is there a difference between LEAP and Traditional on the Force Concept Inventory?
4. Is there a difference between LEAP and Traditional on the Gender Force Concept Inventory?

Sample

In spring 2014, the GFCI was used in lieu of the FCI to assess the 1st semester algebra-based physics course at Tennessee Tech University. The table below shows the number of students in each category who have both pre- and post-test data.

	FCI Fall 2008 — Fall 2013		GFCI Spring 2014	
Pedagogy	Male	Female	Male	Female
Traditional	225	184	33	14
LEAP	178	202	13	15

Pre-analysis

Preliminary data screening for the GFCI posttest measure showed no interaction of pedagogy and gender with the pretest, $F(3, 68) = .301$, $p = .825$, indicating that differences in posttest scores between groups were equally adjusted by pretest performance. The same was true for the FCI, $F(3, 782) = 2.358$, $p = .070$, thus showing that the reported effects of pedagogy, gender, and pretest performance do not overlap. Pre-analysis performed prior to ANOVA & ANCOVA analysis confirmed that other conditions necessary to conduct the appropriate statistics were also met.

Limitations

The results are more tenuous for the GFCI as sample size is so small after elimination of students missing either the pretest or posttest. We plan to acquire additional data this coming academic year.

References

D. Hestenes, M. Wells, & G. Swackhamer, *Phys. Teach. 30*(3), 141–151 (1992).
L. McCullough, & T. Foster, "A gender context for the Force Concept Inventory," In *meeting of the American Association of Physics Teachers, San Diego,* 2001.
L. McCullough, "Gender, Math, and the FCI," Paper presented at Physics Education Research Conference 2002, Boise, Idaho, 2002.
L. McCullough, *J. Int. Womens Stud. 5*(4), 20–30 (2004).
M. Lorenzo, C. Crouch, & E. Mazur, *Am. J. Phys. 74*(2), 118–122 (2006).
L. Kost-Smith, Ph.D. dissertation, University of Colorado Boulder, 2011.
L. Kost, S. Pollock, & N. Finkelstein, *Physical Review Special Topics-Physics Education Research*, *5*(1), 010101(2009).
R. Dietz, R. Pearson, M. Semak, & C. Willis, "Gender bias in the force concept inventory?," In *AIP Conference Proceedings* No. 1413, (AIP, New York, 2012), p. 171.
J. Blue & P. Heller, in *Physics Education Research Conference,* Madison, WI, 2003, pp. 45–48.
R. Hake, *Am. J. Phys. 66*(1), 64–74 (1998).
P. Rosenbaum & D. Rubin, The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55 (1983a).
D. Rubin, "The design *versus* the analysis of observational studies for causal effects: Parallels with the design of randomized trials," *Statistics in Medicine, 26*(1), 20–36 (2007).
L. Rudner & J. Peyton, Consider propensity scores to compare treatments. *Practical Assessment, Research & Evaluation, 11*(9) (2006).

This material is based upon work supported by the National Science Foundation under Grant No. 0737324 and 1245684.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

ANCOVA Results

FCI
Scores on the pre-test, $F(1, 784) = 146.997$, $p = .000$, partial $\eta^2 = .158$, significantly adjusted the post-test and accounts for 15.8% of the variance in post-test scores. After adjusting for the covariance of pre-test, there is sufficient evidence to suggest that there is a difference between post-test means of those enrolled in a LEAP section versus those enrolled in a Traditional section, $F(1, 784) = 264.882$, $p = .000$, partial $\eta^2 = .253$. **Pedagogy is found to be associated with 25.3% of the variance in post-test performance**. The ANCOVA revealed **a small effect of gender on the post-test**, $F(1, 784) = 6.071$, $p = .014$, partial $\eta^2 = .008$, indicating a significant though small difference in post-test scores for males and females.

GFCI
Scores on the pre-test, $F(1, 70) = 15.635$, $p = .000$, partial $\eta^2 = .183$, significantly adjusted the post-test and accounts for 18.3% of the variance in post-test scores. After adjusting for the covariance of pre-test, there is sufficient evidence to suggest that there is a difference between post-test means of those enrolled in a LEAP section versus those enrolled in a Traditional section, $F(1, 70) = 50.251$, $p = .000$, partial $\eta^2 = .418$. **Pedagogy is found to be associated with 41.8% of the variance in post-test performance**. The ANCOVA revealed **no effect of gender on the post-test**, $F(1, 70) = .172$, $p = .680$, partial $\eta^2 = .002$, indicating no significant difference in post-test scores for males and females.

ANOVA Results

FCI
There is sufficient evidence to suggest that there is a difference between the means of the normalized gain of those enrolled in a LEAP section versus those enrolled in a Traditional section, $F(1, 785) = 231.456$, $p = .000$, partial $\eta^2 = .228$. **Pedagogy is found to be associated with 22.8% of the variance in normalized gain**. The ANOVA revealed **a small effect of gender on normalized gain**, $F(1, 785) = 8.080$, $p = .005$, partial $\eta^2 = .010$, indicating a significant though small difference in normalized gains for males and females.

GFCI
There is sufficient evidence to suggest that there is a difference between the means of the normalized gain of those enrolled in a LEAP section versus those enrolled in a Traditional section, $F(1, 71) = 45.542$, $p = .000$, partial $\eta^2 = .391$. **Pedagogy is found to be associated with 39.1% of the variance in normalized gain**. The ANOVA revealed **no effect of gender on normalized gain**, $F(1, 71) = .071$, $p = .791$, partial $\eta^2 = .001$, indicating no significant difference in normalized gains for males and females.

Summary of Findings

FCI	Male				Female			
Pedagogy	N	Pre (SEM)	Post (SEM)	Gain (σ)	N	Pre (SEM)	Post (SEM)	Gain (σ)
Traditional	225	9.8 (0.3)	15.3 (0.4)	0.27 (0.15)	184	6.6 (0.2)	11.9 (0.4)	0.23 (0.14)
LEAP	178	9.0 (0.3)	19.8 (0.4)	0.51 (0.26)	202	6.3 (0.2)	17.9 (0.4)	0.49 (0.24)

Gender-FCI	Male				Female			
Pedagogy	N	Pre (SEM)	Post (SEM)	Gain (σ)	N	Pre (SEM)	Post (SEM)	Gain (σ)
Traditional	33	8.3 (0.5)	10.7 (0.6)	0.11 (0.05)	14	4.9 (0.6)	8.6 (1.2)	0.15 (0.11)
LEAP	13	9.7 (1.0)	19.6 (1.3)	0.49 (0.22)	15	7.1 (0.6)	16.7 (1.4)	0.42 (0.19)

Variable	Method	FCI Partial η^2	GFCI Partial η^2
Pedagogy	Normalized Gain	0.228*	.391*
	Posttest	0.253*	.418*
Pretest	Postest	0.158*	.183*

*All values significant at .05 level.

The table shown above summarizes the results of ANCOVA and ANOVA. The partial η^2 values when multiplied by 100 represents, for example, the percentage of the difference in Normalized Gain that is explained by Pedagogy. The effect of Pedagogy on FCI and GFCI performance was measured by post-test scores as well as by normalized gains. Analysis of normalized gain and post-test showed that Pedagogy was a significant variable in explaining performance on each assessment, accounting for approximately 25% of the differences in FCI performance and 42% of the differences in GFCI performance.

Conclusions & Implications

The findings presented here indicate that, when measured using the GFCI, pedagogy accounted for approximately 42% of the differences in performance. Though caution should be taken in comparing forms of the FCI or in comparing a large sample to one semester of students, there is a level of curiosity surrounding the fact that switching assessments coincided with twice the explained variance. Propensity scoring methods are underway and may substantiate this further.

The non-traditional LEAP pedagogy correlates no differently to the performance of females than males but does correlate to an increase in performance for all. Having high confidence in this finding is especially important when looking at whether or not some pedagogies are more efficacious for a particular gender. The findings support the idea that gender differences are more appropriately addressed by looking at deficits in teaching and deficits in assessments rather than deficits in a particular group of people.

Current Project: Propensity Scoring

Logistic regression is used to generate a propensity score for each student based on the covariates listed to the right as well as scores on FCI & GFCI assessments. To answer questions about the possibility that students self-select the LEAP or traditional sections in a non-random fashion, propensity score matching for an instructor who has taught using both pedagogical approaches will be conducted. Following this pre-analysis, all students from all teachers will be matched using propensity scores to further substantiate prior analysis of the effects of pedagogy and gender on FCI and GFCI performance.

As is done in public health studies where randomized controlled trials are not possible due to ethics associated with forcing a group of people to engage in a behavior that is usually self-selected, establishing a non-random equivalent control group allows us to mimic an experimental design. Drawing a striking example from medical research, we cannot force people to smoke in order to study its relationship with cancer…but we can propensity score them in order to probe the theory that people who smoke are predisposed to smoke. If people cannot be predicted to belong to the smoker group, then those people have randomly assigned themselves to the smoker group. Self-selection without predisposition improves observational studies by mimicking an experiment.

Propensity Scoring Covariates (other than instrument data)	
Gender	Was "AP Physics" taken in high school?
Classification (at time of pretest)	Was "Physics" taken in high school?
Major (at time of pretest)	Was "Physical World Concepts" taken in high school?
Race	Was "Physical Science" taken in high school?
Ethnicity	Was "AP Calculus" taken in high school?
County & state of permanent residence	Was "Calculus" taken in high school?
City & state of High School	Was "PreCalculus" taken in high school?
Name of High School	Highest ACT Composite Score
Highest math course taken (at time of pretest)	Highest ACT Science Reasoning Score
Most recent math course completed (at time of pretest)	Highest ACT Mathematics Score
Time since last math class (at time of pretest)	High School GPA