The Persistence of the Gender Gap in Introductory Physics

Lauren E. Kost, Steven J. Pollock, and Noah D. Finkelstein

Department of Physics, University of Colorado at Boulder, Boulder, Colorado 80309, USA

Abstract. We previously showed[1] that despite teaching with interactive engagement techniques, the gap in performance between males and females on conceptual learning surveys persisted from pre- to posttest, at our institution. Such findings were counter to previously published work[2]. Our current work analyzes factors that may influence the observed gender gap in our courses. Posttest conceptual assessment data are modeled using both multiple regression and logistic regression analyses to estimate the gender gap in posttest scores after controlling for background factors that vary by gender. We find that at our institution the gender gap persists in interactive physics classes, but is largely due to differences in physics and math preparation and incoming attitudes and beliefs.

Keywords: gender, conceptual learning, introductory physics, physics education research
PACS: 01.40.Fk, 01.40.G-, 01.40.gb, 01.75.+m

INTRODUCTION

Previously, we have observed a difference of about 10 percentage points in both the pre- and posttest performance of males and females on the Force and Motion Concept Evaluation[3] at our institution, averaging over seven semesters of the introductory physics course[1]. The posttest gender gap exists in both partially interactive and fully interactive courses, though the size of the posttest gap varies from semester to semester (from 5% to 15%). Previous work[4] suggested that differences in student background and preparation may contribute to the persistence of the gender gap. When students were grouped according to their pretest scores, the average posttest scores of males and females in each group were not significantly different ($p>0.1$), though males did consistently score higher in each group than females.

In order to identify factors that may contribute to the persistence of the gender gap, we continue to analyze student background factors that may influence performance on the conceptual learning survey. We use two types of regression analysis (multiple regression and logistic regression) to estimate the impact of student background factors on conceptual test performance. We find that the gender gap in conceptual posttest scores is substantially accounted for once prior physics and mathematics understanding and incoming attitudes and beliefs are taken into account.

METHODS

The data in the following studies were collected from six offerings (fall 2004 to spring 2007) of the first semester, calculus-based mechanics course at the University of Colorado (CU) and represent over 3000 students. These are large-enrollment courses with 400 to 600 students. Each semester was taught by a different instructor, and all six instructors were male. All six classes used interactive engagement techniques, some to a higher degree than others. Each of the six classes employed student discussions around ConcepTests[5] in lecture, online homework systems[6], and voluntary help-room sessions on problem-solving homework. Three of the six classes used *Tutorials in Introductory Physics*[7] and Learning Assistants[8] during a one-hour per week recitation, while the remaining three classes held more traditional recitation sections. There is no laboratory associated with this course.

The student population in the introductory course is about 25% female, about 50% engineering majors, with over 80% white. Only about 6% of the students who enroll in introductory physics are declared physics majors, and 66% of those students change majors after the first semester.

Of primary interest in this study is to what degree differences in the backgrounds of male and female students contribute to the observed gender gap. To that end, data have been gathered on students' background knowledge and preparation for college physics, as well as their conceptual performance in

CP1064, *2008 Physics Education Research Conference*, edited by C. Henderson, M. Sabella, and L. Hsu

the course, shown in Table 1. Except for FMCE and CLASS (attitudes and beliefs)[10] results, all data were collected from the Office of Planning, Budget, and Analysis at CU, rather than self-reported.

TABLE 1. Data collected and the variables they measure. The * indicates that the males' average is significantly higher than the females' average.

Variables	Data Collected
Physics conceptual understanding at the end of the semester	FMCE posttest*
Prior physics conceptual understanding	FMCE pretest*
Prior academic achievement	High school GPA
Prior mathematics understanding	Combined Math Score* [9] (SAT-Math*, ACT-Math*, 2 CU placement exams)
Course preparation for college physics	Yrs. high school physics* Yrs. high school calculus
Prior attitudes and beliefs about physics and about learning physics	CLASS pretest*
Demographics	Gender Class Standing Ethnicity Declared Major

The assessments used in this study only measure student *performance* on these instruments – however we use them as a proxy measurement of student understanding and attitudes and beliefs upon entry and exit. We recognize these instruments may be measuring more (e.g. test-taking ability), and may differ by gender. The work of several researchers suggests that differences by gender on conceptual instruments may indicate something other than differences in physics understanding.[11-13] While these studies question the validity of these measures, we note: a) we are using the standard measures that have been adopted by the community and b) we are analyzing *shifts* on these instruments, which allows us to normalize students against themselves.

Not all of the above data are available for all students. Of the 3,205 students that enrolled in introductory physics during the semesters in this study, data on all variables was available for only 1,027 students. These 1,027 students make up the sample used in the analysis. The students in the sample are distributed about equally among each of the six semesters. It is important to point out that the percentage of females in this sample is 29% (higher than in the population of all students) and the average course grade of students in the sample is significantly higher than that of students not in the sample ($\langle GPA_{sample}\rangle=2.9$, $\langle GPA_{not\text{-}sample}\rangle=2.4$, $p<0.01$). While not completely representative, this subsample suggests that we are underestimating the gender gap of all students.

RESULTS: Multiple Regression

Using the background and demographic variables in Table 1, we model the posttest scores using multiple regression[14]. The prediction equation (Eq. 1) is used to estimate the difference in posttest scores between a male and female for background variables held equal. However, due to the nature of our data, notably non-Gaussian distributions[15], we cannot reliably interpret the statistical significance of the results, as they are likely to be biased. We do interpret the results of the regression analysis qualitatively, noting which trends exist. In this way, we can get a sense of whether the gender gap can be accounted for by factors other than gender.

The posttest scores are modeled according to the equation,

$$FMCEPOST = a + b_1 \times FEMALE + \sum_{k=2}^{N} b_k \times VAR_k \quad (1)$$

where *FMCEPOST* is the posttest score, *FEMALE* is 1 for females (0 for males), and VAR_k are the other background variables included in the model and any cross terms between *FEMALE* and other variables. The multiple regression analysis estimates the coefficients for each term, i.e. the b_k's .

The results are shown in Table 2, which contains the coefficient estimates as well as the model-level statistics. The Multiple R^2 (the fraction of variation in posttest scores accounted for by the variation of the variables in the model) is 0.44.

TABLE 2. Coefficient estimates and model-level statistics for the final regression model.

Model-Level Statistics	
Multiple R-squared	0.44
F statistic p value	<0.001
Predictors	**b**
Intercept	32.9
Female	-9.2
FMCE Pretest	0.6
Combined Math Score	7.2
CLASS Pretest	0.3
2004 Fall Semester	1.3
2005 Spring Semester	-5.6
2005 Fall Semester	-8.7
2006 Spring Semester	-2.9
2006 Fall Semester	-0.9
Female*FMCE Pretest	0.2

While all variables in Table 1 were examined, the final best-fit model only includes variables that control for prior physics and mathematics understanding, prior attitudes and beliefs, the semester the student was enrolled, and an interaction term between *FEMALE* and the pretest score. Although we have no further information about specific aspects of each semester that could contribute to the differences, the size of the semester variables demonstrates that there are considerable differences by semester, even after other prior factors are accounted for. The interaction term is the product of the two variables, *FEMALE* and *FMCEPRE*, and allows the slope of the pretest variable to differ for males and females. This term suggests that the pretest score differently predicts the posttest score for males and females.

This model allows us to estimate the difference between a male's and a female's posttest scores, controlling for several other factors. The difference in posttest scores is given by

$$M - F = 9.2 - 0.2 \times FMCEPRE. \qquad (2)$$

The average pretest score for this sample of students is 30.3. The gender difference for a male and a female, each with the average pretest score (all other variables equal) is 3.2 points. This is a substantial reduction from the 10.7 point difference observed by subtracting the average male and female posttest scores. The effect size[16] is reduced from 0.39, when no background variables were controlled, to 0.11, when measures of student background are controlled for. Using multiple regression analysis, 7.5 points of the 10.7 point gender difference, or 70%, can be attributed to differences in male and female backgrounds.

While the final model includes many variables that one might suspect would influence posttest scores, there are several variables that are not included. Variables that were found to not substantially impact the posttest score or gender gap beyond those variables in the final model were: years of high school physics, years of high school calculus, high school GPA, declared major, ethnicity, and interaction terms between *FEMALE* and all other variables. This is not to argue that these factors do not matter, but rather that either they are represented by other variables (e.g. pretest score makes years of high school physics less significant) or the variance among these variables was too small (e.g. ethnicity).

RESULTS: Logistic Regression

Given the distribution of data, the multiple regression analysis only allows for qualitative interpretations. One way to analyze the data statistically, given the non-normal distribution of data, is to use logistic regression analysis[17]. While using this method allows us to make statistical claims, we lose the ability to predict students' actual posttest score and can only predict whether they will score above some threshold. To perform the analysis the FMCE posttest is converted into a categorical variable that equals 1 if the posttest score is above 60% and 0 if the posttest score is below 60% [18].

We observe that 64% of the males and 49% of the females score above 60% on the FMCE posttest. These percentages are significantly different (via χ^2, $p<0.01$). This difference is the gender gap that we are concerned with in the logistic regression analysis.

The posttest data are modeled according to the equation,

$$\ln(odds(FMCEPOST > 60\%)) = a + b_1 \times FEMALE + \sum_{k=2}^{N} b_k \times VAR_k \qquad (3)$$

where odds(X) is defined as the probability of X divided by the probability of not-X [17]. Given that the gender gap in this analysis is the difference in odds of scoring above 60% for males and females, we are interested in whether the difference in odds can be explained by factors other than gender. The odds for a male and a female, all other variables being equal, are related according to the equation,[1]

$$odds_F = g \times odds_M \qquad (4)$$

where $g = e^{b_1}$. The logistic regression analysis estimates the coefficients of each variable, which then allows a prediction of each student's odds of scoring above 60%. We are ultimately interested in whether the *FEMALE* variable is significantly different from zero (as indicated by the p value) and whether *g* is less than, greater than, or equal to 1. The final regression model is presented in Table 3.

TABLE 3. Coefficient estimates and model-level statistics for the final logistic regression model.

Model-Level Statistics		
Pseudo R-Squared (Nagelkerke's)		0.45
Likelihood Ratio p value		<0.001
Predictors	**b**	**p value**
Intercept	-2.93	<0.001
Female	-0.23	0.19
FMCE Pretest	0.08	<0.001
Combined Math Score	0.55	<0.001
CLASS Pretest	0.02	<0.001
2004 Fall Semester	0.14	0.59
2005 Spring Semester	-0.68	0.01
2005 Fall Semester	-0.75	0.01
2006 Spring Semester	-0.10	0.71
2006 Fall Semester	-0.06	0.81

[1] $odds_F = odds(FMCEPOST_F > 60\%)$

The final model here is the same as the multiple regression model (Table 2 and Eqn. 1) except the logistic regression model does not include the interaction term (as it was not significant). Before any background variables are taken into account, the factor relating the odds of males and females is g = 0.5. The final estimate of *g*, all other variables held constant, is

$$odds_F = 0.8 \times odds_M . \qquad (5)$$

g is just slightly smaller than 1 (and not statistically different from 1), meaning that the odds for males and females are statistically equal. Using logistic regression analysis, we find that about 57% of the gender gap in odds is accounted for by prior physics and math knowledge and prior attitudes and beliefs.

DISCUSSION AND CONCLUSIONS

Both the multiple regression and the logistic regression analyses confirm that a substantial fraction of the gender gap can be accounted for by differences in males' and females' preparation and background coming into the introductory course, and is not explicitly due to student gender. Other researchers have found that these same factors, prior physics experience[19] and math understanding[20], influence conceptual performance and course grade. Though, we point out that only about 45% of the variance in our posttest scores is explained by the final models. Other factors that could influence the posttest score, such as student motivation, self-efficacy, study habits, socioeconomic status, specific aspects of curriculum implementation, and testing bias, have not been included.

In neither analysis has the gender gap been completely accounted for. Using multiple regression, 70% of the gender gap in posttest scores is accounted for by background, while using logistic regression, 57% of the gender gap in odds is accounted for by background. Both analyses suggest that even after accounting for background differences between males and females, a small gender gap remains.

While the classes studied are introductory courses with no expectation of prior knowledge of physics (there are no prerequisites for this course), those students who arrive to the class with greater background knowledge (higher pretest scores) are more likely to achieve high posttest scores. Such an arrangement of a class plays to certain student backgrounds and when those backgrounds are correlated with particular demographic groups, it demonstrates bias. Recognizing that student preparation in physics or mathematics is a means by which this bias is propagated allows us, as researchers and educators, to proactively address the challenges of the gender gap in physics.

ACKNOWLEDGMENTS

This work is supported by PhysTEC (APS/AIP/AAPT), NSF CCLI (DUE0410744) and NSF CAREER (REC0448176). Thanks to the PER at Colorado group and the faculty that were involved. Thanks to D. Briggs and J. Watkins for help with statistical analyses.

REFERENCES AND NOTES

1. S.J. Pollock et al. *Phys. Rev. ST PER* **3**, 010107 (2007).
2. M. Lorenzo et al. *Am. J. Phys*. **74**, 118 (2006).
3. R.K. Thornton and D.R. Sokoloff, *Am. J. Phys*. **66**, 228, (1998).
4. L.E. Kost et al, "Investigating the Source of the Gender Gap in Introductory Physics", *2007 Physics Education Research Conference*, AIP Conf. Proceedings 951, AIP, Melville, NY, 2007, pp. 137-140.
5. E. Mazur, *Peer Instruction: A User's Manual*, Upper Saddle River, NJ: Prentice-Hall, 1997.
6. CAPA, http://www.lon-capa.org/; Mastering Physics, http://www.masteringphysics.com/.
7. L.C. McDermott and P.S. Schaffer, *Tutorials in Introductory Physics*, Upper Saddle River, NJ: Prentice-Hall, 2002.
8. V. Otero et al. *Science* **313**, 445 (2006).
9. Exams are given to students before their freshman year at CU to place them in the appropriate math course. The combined math score was calculated by first converting the scores on each test to z-scores, $z_i=(S_i - \langle S \rangle)/SD_S$, and then averaging the normalized scores.
10. W. Adams et al., *Phys. Rev. ST PER* **2**, 010101 (2006).
11. L. McCullough, *J. Int. Women's Stud*. **5**, 20 (2004).
12. T.L. McCaskey et al., "Effects in assessment caused by splits between belief and understanding", *2003 Physics Education Research Conference*, AIP Conf. Proceedings 720, AIP, Melville, NY, 2003, pp. 37-40.
13. E. Hazel et al. *Int. J. Sci. Ed*. **19**, 381 (1997).
14. Lomax, R.G., *An Introduction to Statistical Concepts for Education and Behavioral Sciences*, Mahwah, NJ: Lawrence Erlbaum Associates, 2001.
15. Because of ceiling effects, heteroskedasticity, non-normal data, and non-random sampling, our data do not meet the assumptions of multiple regression that allow for unbiased interpretation of statistical significance.
16. Effect size is $(\langle posttest \rangle_M - \langle posttest \rangle_F)/SD_{posttest}$, where $SD_{posttest}$ is the standard deviation of the distribution of posttest scores for all students.
17. Mendenhall, W. and Sincich, T., *A Second Course in Statistics: Regression Analysis*, Upper Saddle River, NJ: Pearson Education Inc., 2003.
18. The analysis was run for several threshold values and also for several numbers of posttest categories. The results were similar in all cases.
19. Z. Hazari, et al., *Sci. Ed*. **91**, 847 (2007).
20. D.E. Meltzer, *Am. J. Phys*. **70**, 1259 (2002).