Characterizing 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
(Received 26 June 2008; published 8 January 2009)

Previous research [S. J. Pollock *et al.*, Phys. Rev. ST Phys. Educ. Res. **3**, 1 (2007)] showed that despite the use of interactive engagement techniques, the gap in performance between males and females on a conceptual learning survey persisted from pretest to post-test at the University of Colorado at Boulder. Such findings were counter to previously published work [M. Lorenzo *et al.*, Am. J. Phys. **74**, 118 (2006)]. This study begins by identifying a variety of other gender differences. There is a small but significant difference in the course grades of males and females. Males and females have significantly different prior understandings of physics and mathematics. Females are less likely to take high school physics than males, although they are equally likely to take high school calculus. Males and females also differ in their incoming attitudes and beliefs about physics. This collection of background factors is analyzed to determine the extent to which each factor correlates with performance on a conceptual post-test and with gender. Binned by quintiles, we observe that males and females with similar pretest scores do not have significantly different post-test scores ($p > 0.2$). The post-test data are then modeled using two regression models (multiple regression and logistic regression) to estimate the gender gap in post-test scores after controlling for these important prior factors. These prior factors account for about 70% of the observed gender gap. The results indicate that the gender gap exists in interactive physics classes at our institution but is largely associated with differences in previous physics and math knowledge and incoming attitudes and beliefs.

DOI: 10.1103/PhysRevSTPER.5.010101 PACS number(s): 01.40.Fk, 01.40.G−, 01.40.gb

I. INTRODUCTION AND BACKGROUND

A recent American Institute of Physics (2005) report found that females earned 22% of all bachelor's degrees and 18% of all doctoral degrees in physics.[1] At the University of Colorado (CU), females make up only 25% of the students who enroll in introductory physics and about 15% of the physics majors. Not only is there a gender gap in participation, but there is also a gender gap in performance. Previous studies at CU, and elsewhere, have identified differences in males' and females' performances on surveys of conceptual physics.[2–4] This under-representation and underperformance of females in physics is cause for concern and has led to a variety of studies on the source of the gender gaps in college physics.[1,5,6] To further understand the gender gap at CU, we continue to look at student performance in introductory physics. As this is the first college physics course that students take, and sometimes their first encounter with physics, it serves as an important first step toward pursuing a physics degree. In a previous study,[7] 45% of students reported that their interest in physics decreased over the course of the first semester of introductory physics at CU. Of those that gave a reason for their decreased interest, one third indicated their personal success (or failure). Student success in this course can impact whether students continue in the major. Understanding the factors that do and do not promote student learning in this course is crucial to understanding the gender gap and finding ways to eliminate it. The current work identifies several factors that are correlated with student post-test performance on a conceptual learning survey and then estimates the extent to which these factors can account for the observed gender gap. The goal of this work is to determine how much of the observed gender gap can be attributed to factors other than gender explicitly.

Prior research has consistently demonstrated the benefits of using interactive engagement techniques during instruction.[8–10] Hake's[8] survey of traditional and interactive courses found that interactive engagement courses had average normalized learning gains,

$$\langle g \rangle = \frac{\langle \text{post} \rangle - \langle \text{pre} \rangle}{100 - \langle \text{pre} \rangle}$$

almost two standard deviations higher than average gains in traditional courses. Learning gains at CU reach as high as $\langle g \rangle = 0.64$ for reformed courses, which make use of several research-based interactive techniques.[11,12] While the use of interactive engagement techniques has been shown to facilitate learning for both males and females, some research has suggested that females may benefit more than males.[13,14] Researchers at Harvard University found that a preinstruction gender gap was eliminated over the course of an interactive and engaging introductory physics course.[3] Both males and females had significant gains in the interactive course, but females had slightly larger gains, resulting in males and females having about equal post-test scores.

At CU,[4] and elsewhere,[2,15] the gender gap has persisted despite the use of interactive engagement techniques. Figure 1 shows the pretest and post-test gender gaps for partially interactive and fully interactive courses (described more below).[16] The post-test gender gap is smaller for fully interactive courses than for partially interactive courses, but it has not been eliminated.[17] Furthermore, we demonstrate (below) the gender gap varies from semester to semester. Similar gender gaps are found when looking at normalized learning gains. For both partially and fully interactive courses males have a higher average normalized learning gain than females,

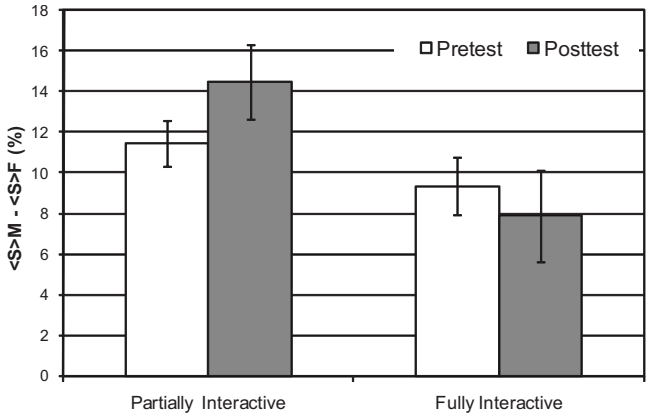

FIG. 1. Pretest and post-test gender gaps ($\langle S\rangle_M - \langle S\rangle_F$) for partially and fully interactive courses. Student performance on the FMCE is averaged over three semesters (partially interactive) and four semesters (fully interactive). Error bars represent the standard errors of the mean. There is no statistically significant shift in the gender gap for either partially or fully interactive courses.

although the difference between males and females is smaller for fully interactive courses.[4]

The fact that there is a gender gap in the pretest scores (before any instruction) suggests that there are differences in preparation between male and female students when they enter the introductory physics course, and we suspect that these differences may contribute to the persistence of the gender gap. Several researchers have investigated the factors that influence student performance in introductory physics. Hazari *et al.*[6] found that mathematics preparation was a significant predictor of students' college physics grade. Others have found an influence of high school physics experience on college physics performance. Sadler and Tai[18] find that taking a high school physics course is positively related to college physics course grade, even when controlling for students' self-reported academic and demographic background. Furthermore, the pedagogy of the high school class is related to a student's college performance. Students who take high school classes that cover fewer topics in more depth have higher grades in a calculus-based college course than students whose high school classes cover more topics in less depth (the difference is almost a full letter grade).[19] Affective factors, such as father's encouragement and family beliefs about science, have also been shown to influence student performance.[6] These studies have all focused on students' course grades as the measure of student performance. Less work has been done on the factors that influence student performance on research-based conceptual learning surveys. As stated above, the curriculum and level of engagement in the course influences student conceptual learning.[2–4] Meltzer[20] found correlations (Pearson's product-moment correlation coefficient of $0.1 < r < 0.5$) between a students' score on a preinstruction math skills test and their normalized gain on an electricity and magnetism conceptual survey for four algebra-based physics courses.

While we can look at a variety of measures to assess student performance (post-test, normalized gain, course grades, etc.), we focus on the post-test as an objective measure of what students know at the end of the semester. While normalized gain can also be used, we opt to use the post-test as a measure of students' physics knowledge after a semester of instruction, rather than their gain in knowledge over the course of a semester. Course grades are subjective and measure more than just performance on a single instrument. Prior work has mostly focused on subjective measures of student performance and the self-reported backgrounds of students. The current study aims to identify several prior factors that influence student performance on a research-based mechanics conceptual learning instrument using data collected from university applications for students in an introductory calculus-based physics course.

In this paper, we seek to characterize the gender gap in introductory physics and those factors that are correlated with the differential performance at our institution. Identifying these factors is a first step in clarifying the mechanisms by which the gender gap is established and will lead to future work that guides interventions that address this disparity. We address the following research questions: (1) On what measures do we observe differences by gender in the introductory physics course, for example, conceptual learning, components of the course grade, attitudes and beliefs, and prior knowledge and preparation? (2) Are measures of background correlated with student performance in the course (as measured by a conceptual learning survey) and correlated with gender? (3) To what extent do differences in males' and females' backgrounds contribute to the persistence of the gender gap in introductory physics courses? We find that there are several aspects of the introductory course in which we identify gender differences, including males' and females' course grades, prior physics and mathematics understanding, and their attitudes and beliefs about physics. When these differences in prior understanding and attitudes are controlled, the gender gap is substantially accounted for.

II. RESEARCH METHODS

The data in the following studies were collected from seven offerings (from Spring 2004 to Spring 2007) of the first-semester calculus-based introductory mechanics course at the University of Colorado. These are large-enrollment courses that typically have 400–600 students. Each semester was taught by a different instructor, and all seven instructors were male. All seven classes used interactive engagement (IE) techniques, some to a higher degree than others. Each of the seven classes employed student discussions around ConcepTests[11] in lecture, online homework systems,[21] and voluntary help-room sessions on problem-solving homework. Four of the seven classes used *Tutorials in Introductory Physics*[12] and Learning Assistants[22] during a 1 h/week recitation, while the remaining three classes held more traditional recitation sections. There is no laboratory associated with this course. A more detailed description of the course structure can be found in previous work.[23] We categorize the three classes that held traditional recitation sections as IE 1 (partially interactive) and the four classes that used *Tutorials* during recitation sections as IE 2 (fully interactive). Our definitions of IE 1 and IE 2 classes are similar, but not identical, to the definitions used in prior studies.[3,4,24]

TABLE I. Frequencies for gender, student declared major, and ethnicity for all students in the study, that is, students who enrolled in introductory physics between Spring 2004 and Spring 2007.

Gender (N=3728)	%		
Male	75.8		
Female	24.2		
Major (N=3728)	%	% of males	% of females
Physics	5.5	5.6	5.2
Engineering	51.8	55.6	39.9
Other science	18.1	14.1	30.5
Nonscience	9.9	8.4	14.6
Undeclared or other	14.8	16.4	9.9
Ethnicity (N=3514)	%	% of males	% of females
Asian	8.9	8.1	11.5
African American	1.3	1.3	1.1
Hispanic	6.1	6.1	6.3
Native American	0.8	0.9	0.7
White	81.1	81.9	78.7
Foreign	1.7	1.7	1.8

The student population in the introductory course is about one-quarter female. About half of the students are engineering majors and about 20% are other science majors. Only about 6% of the students who enroll in introductory physics are declared physics majors.[25] There are some differences in the distributions of student major for males and females, but the same percentage of males and of females are physics majors, as seen in Table I. Females are less likely than males to be engineering majors but about twice as likely as males to be other science or nonscience majors. Over 80% of the students are white, about 10% are Asian, and about 8% are African American, Hispanic, or native American. There are only small differences in the distributions of ethnicity by gender. These frequencies are presented in Table I.

Of primary interest in this study is to what degree males and females differ on measures of background and preparation and to what degree these differences contribute to the observed gender gap. Conceptual performance, as measured by the Force and Motion Concept Evaluation (FMCE),[26] serves as the focus of the study. The FMCE post-test score for each student is used as a measure of the student's conceptual knowledge of physics at the end of the semester. Only students with matched pretest and post-test data are included ($N=2099$). Additional evaluation of student performance in the course is captured by homework, exam, participation, and course grades, which were collected from the instructor in each course.

Data have been gathered[27] on students' background knowledge and their preparation for college physics. Prior academic performance is captured by students' high school grade point average (GPA), while the FMCE pretest is used to measure students' prior conceptual understanding of physics. Four mathematics tests are combined to form a measure of students' prior knowledge of mathematics. The four tests include the math portion of the Scholastic Aptitude Test (SAT-math), the math portion of the American College Test (ACT-math), and two diagnostic exams that are given to students before their freshman year at CU. One diagnostic exam (denoted APPM test), is given through the Applied Mathematics Department to students in the School of Engineering. The second exam (denoted ASMATH test), is given through the Mathematics Department to students in the School of Arts and Sciences. Both diagnostic exams are used to help place students in the appropriate math course and do not count toward any course grade. Scores on each of the four tests were similarly correlated with the FMCE post-test ($0.3<r<0.4$) and were also highly correlated ($0.5<r<0.7$) with each other. To get a measure of prior math knowledge for almost every student and to avoid having multiple variables that contained the same information, the scores on the four tests were combined. The scores for each test were first normalized (converted to z scores[28]). Each student's normalized scores were averaged to get a combined measure of mathematics knowledge. Each student's combined math score is a composite of whichever of the four tests that the student took. Student course preparation for college physics is measured by how many years of high school physics and calculus a student had taken. Data were not available on the grade that students received in their high school courses.

In addition to students' prior content knowledge, data were also collected on their attitudes and beliefs about physics and about learning physics. Attitudes and beliefs are measured by the Colorado Learning Attitudes about Science Survey (CLASS).[29] The CLASS questions are classified into eight categories of student beliefs. The survey is made up of 42 statements and students respond on a Likert-type scale. Each response is coded favorable, neutral, or unfavorable based on whether the response agrees or disagrees with the expert response. Students are then given a percent favorable and a percent unfavorable score on each category. Pretest scores on each category are used as measures of students' incoming beliefs. Post-test scores and shifts (post−pre) are used as measures of students' attitudes and beliefs at the end of the semester and to measure change in attitudes and beliefs, respectively. Throughout these analyses only the percent favorable scores will be considered.

We note that the several assessments used throughout the study only measure student *performance* on these instruments—however we use them as a proxy measurement of student understanding and actual attitudes and beliefs upon entry and exit. We recognize that these instruments may be measuring more, such as test taking ability, and may differ by gender. In particular, McCullough[30] found that, by changing the context of questions on the Force Concept Inventory (FCI) (Ref. 31) to gender neutral or female contexts, male and female students responded differently. Other researchers identified differences in how students responded on the FCI when asked to mark the answer that they believed and the answer they thought scientists would give. Females answered differently in each case more often than males.[32] Still others have pointed out that the format of questions that are typically asked in physics classes (multiple choice questions) may disadvantage females.[33] While these studies ques-

TABLE II. Average course grades for males and females who did and did not take the FMCE. Course grades are on a 0–4.0 scale (Ref. 34).

	Males			Females			Differences	
	N	Mean	SD	*N*	Mean	SD	*M*−*F*	*p* value
Students without FMCE	1152	2.14	1.2	315	1.89	1.1	0.25	0.001
Students with FMCE	1563	2.82	0.8	533	2.74	0.8	0.08	0.086

tion the validity of these instruments, we note that (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.

The FMCE is administered the first and last weeks of classes during recitation, and only those students that attend both weeks take the pre- and post-FMCE. As a result, we explore the possibility of sampling bias. Of the 3728 students who took introductory physics during the semesters included in this study, 2099 students (56%) took both the pre- and post-FMCE. Comparing the populations that did and did not take the FMCE, we find that females were more likely to take the FMCE than males: the sample that did take the FMCE is 25% female, and the sample that did not take the FMCE is 21% female. The course grades (on a scale from 0.0 to 4.0) for males and females in each group are shown in Table II. Not only are the average course grades of students who take the FMCE higher, but the gender gap in course grades for this group is smaller than for those that do not take the FMCE. By focusing on the FMCE as a measure of learning we limit the sample of students included in the analysis and exclude primarily those with lower course grades. Also, the smaller gender gap in course grades among those that take the FMCE suggests that we may be underestimating the gender gap.

III. RESULTS: IDENTIFYING DIFFERENCES BY GENDER

A. College course performance differences

Figure 2 presents the pretest and post-test gender gaps for each semester included in the present study. A different instructor lectured during each of the seven semesters. Counter to previous findings,[3] we find that the size of the post-test gender gap is not independent of instructor. There is some consistency within the IE 1 and IE 2 courses. In all three IE 1 courses the gender gap increased (although not significantly) from pretest to post-test. In three of the four IE 2 courses, the gender gap decreased (although not significantly). But there is one IE 2 class (semester E) in which the gender gap increased. These findings suggest that the implementation of a fully interactive curriculum alone is not enough to eliminate or even reduce the gender gap. It appears that the manner in which courses are implemented is significant—which is the subject of current study.[35] Furthermore, the way in which the curriculum is enacted may appear to impact the gender gap; however, we find (below) in these cases that differences in the gender gap from semester to semester can largely be accounted for by background differences of the students.

As reported above, females have lower normalized learning gains than males, meaning that females learn a smaller percentage of what they did not already know coming into the introductory course. In addition to looking at normalized learning gain, we also look at male and female average absolute gain (G=post−pre). The average absolute gain for both males and females is statistically significantly higher in IE 2 courses than in IE 1 courses, but in neither pedagogical approach is the difference between average absolute gains for males and females significantly different ($p>0.1$ and unless stated otherwise, all p values are calculated via two-tailed t-test). In IE 1 courses, the average gain of males and females is G_M=32% and G_F=29%. In IE 2 courses, the average gain of males and females is G_M=37% and G_F=39%. In three of the four IE 2 courses females had larger average absolute gains than males, while in all three IE 1 courses

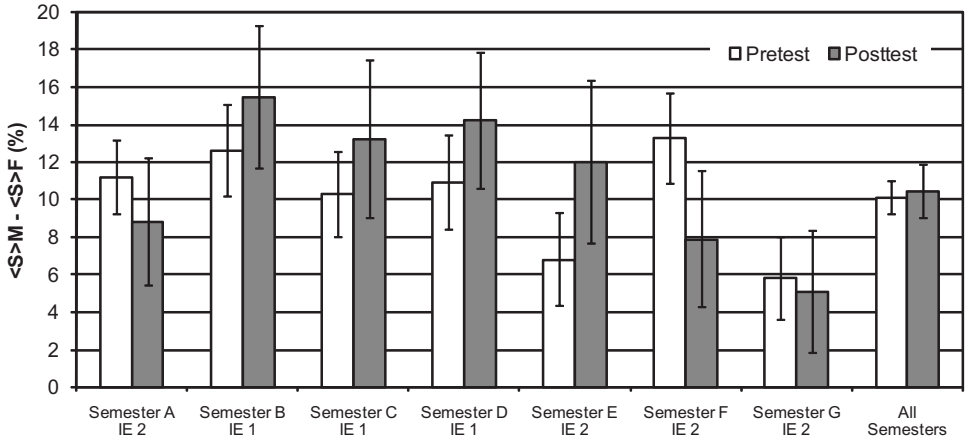

FIG. 2. Gender gaps ($\langle S\rangle_M - \langle S\rangle_F$) in each semester. IE 1 and IE 2 indicate partially and fully interactive courses, respectively. The data represent seven different instructors and over 2000 students. Error bars represent the standard errors of the mean.

TABLE III. Analysis of students' course grades. Each column contains the difference between the average scores for males and females ($\langle S\rangle_M - \langle S\rangle_F$). Error (shown in parentheses) is computed from the standard errors of the mean for males and females added in quadrature. The asterisk (*) indicates that the difference is statistically significant at the $p < 0.05$ level.

	Participation (%)	Homework (%)	Exams (%)	Course GPA (4 pt. scale)
Semester A	−6.6(1.6)*	−7.6(1.9)*	4.9(1.6)*	0.04(0.10)
Semester B		−5.0(2.0)*	3.4(1.2)*	0.11(0.11)
Semester C		−4.8(1.8)*	3.7(1.6)*	0.10(0.11)
Semester D		−5.0(2.0)*	6.3(1.5)*	0.10(0.10)
Semester E	−4.9(1.8)*	–2.9(1.9)	5.2(1.5)*	0.17(0.11)
Semester F	−8.1(1.8)*	−2.0(2.0)	4.8(1.6)*	0.15(0.12)
Semester G	−3.0(1.6)	−3.0(2.0)	3.3(1.4)*	0.06(0.11)
Average	−5.6(0.9)*	−4.5(0.8)*	4.5(0.6)*	0.11(0.04)*

females had lower average absolute gains than males, although none of the differences was significant. Some have suggested that absolute gain may be a more appropriate way to assess learning.[36] We observe that in terms of absolute learning gain, there is no statistically significant gender difference in any individual course or across all courses.

Course grades were examined to determine if males and females perform differently on course grades or any components of the course grades. For each of the seven semesters of the mechanics course males' and females' scores are averaged on homework, participation, exams, and total course grade. In all of the introductory courses exams make up 60%–65% of the course grade, homework counts for 25%–35%, and participation makes up the remainder. The difference between the average male and female's scores in each component ($\langle S\rangle_M - \langle S\rangle_F$) is calculated for each class. These differences for each semester, along with the average differences across all semesters, are shown in Table III. For several courses the participation grade was included in the homework grade and could not be extracted.

There was no significant gender difference in total course grade in any individual course of the seven semesters in the study. Males outscore females by about 5 points on exams and females outscore males by about 5 points each on homework and participation. These differences offset one another and result in course grades that are not significantly different. Because of the consistent gender gap observed from semester to semester, we find that the difference in overall course grades of males and females is statistically significant when averaging over all seven semesters.

In addition to looking at performance, we can also explore how the attitudes and beliefs of males and females change over the course of the semester and whether there are any gender differences. Developers of the CLASS identified gender differences on almost half of the statements and found that, on average, females were less expertlike in their beliefs than males at the end of an introductory calculus-based physics course.[29] Here, we present the average shifts (post−pre) for males and females overall and in each category for six semesters of the introductory calculus-based physics course.[37] Shifts indicate how much students' attitudes and beliefs have changed from the beginning to the end of the semester. As Fig. 3 shows, all of the shifts are negative, indicating that both males and females shift toward less expertlike beliefs about physics over the course of the introductory physics class.[38] In addition, females have *more* negative shifts than males overall and in each category. The difference in shifts is significant ($p < 0.05$) for the three problem-solving and two conceptual categories. Females have pretest scores that are similar to or lower than males' pretest scores in each category (as shown in parentheses in Fig. 3). The larger negative shifts result in an increase in the gender gap in CLASS scores from pretest to post-test in all categories. The introductory physics course appears to be influencing the attitudes and beliefs of males and females differently.

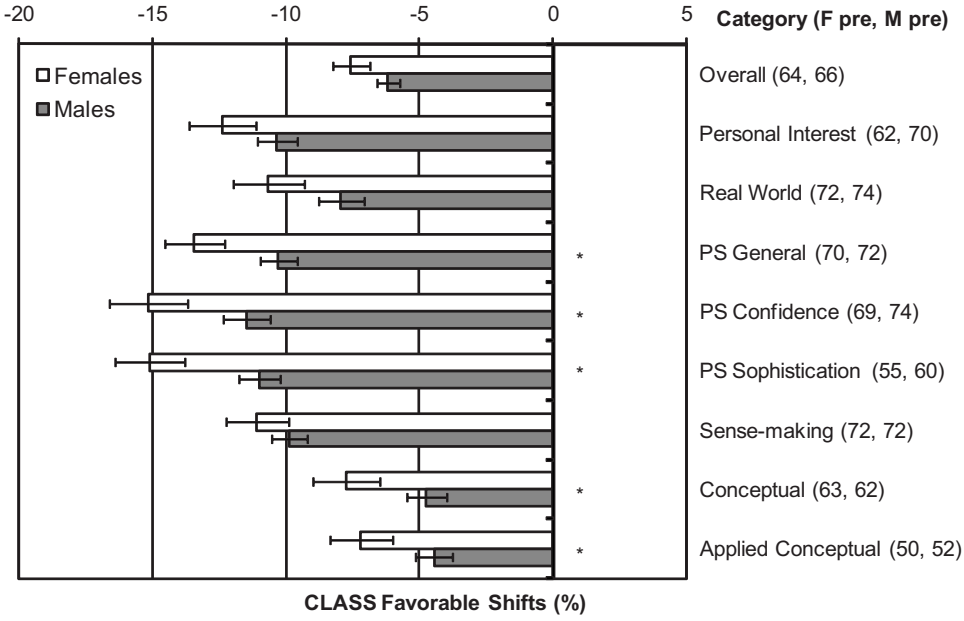

FIG. 3. Average shifts (post−pre) for males and females on each of the CLASS categories. Note that all shifts are negative, meaning both male and female students shift toward less expertlike attitudes and beliefs about physics. The asterisk (*) indicates that the difference in shifts for males and females is significant ($p < 0.05$). Values in parentheses are female and male average pretest scores. Females have more negative shifts in each category than males.

TABLE IV. Male and female average values for all background variables that were collected. The range of possible scores for each variable is shown in parentheses. The effect size is calculated as ES=($\langle S\rangle_M - \langle S\rangle_F$)/SD, where the SD for all students is used. Significant differences exist between males and females on almost all of the background variables.

	Males			Females			Differences		
	N	Mean	SD	N	Mean	SD	$M-F$	Effect size	p value
High school GPA (0–4)	2712	3.55	0.4	869	3.74	0.4	−0.19	0.47	<0.001
Yrs. high school physics	2463	1.08	0.6	758	0.95	0.5	0.13	0.24	<0.001
Yrs. high school calculus	2445	0.78	0.6	748	0.81	0.6	−0.03	0.05	0.212
SAT-math (200–800)	1903	645	65	629	623	75	22	0.33	<0.001
ACT-math (1–36)	2130	28.1	4	732	27.6	4	0.5	0.14	0.001
APPM test (0–30)	1189	22.0	4	255	21.9	5	0.1	0.04	0.625
ASMATH test (0–30)	349	17.1	2	128	17.3	2	−0.2	0.06	0.599
Math combined (z score)	2744	0.024	0.9	869	−0.184	1	0.21	0.23	<0.001
CLASS pretest (0–100)	1380	65.7	16	522	63.6	16	2.1	0.13	0.012
FMCE pretest (0–100)	1566	32.2	21	533	22.0	16	10.2	0.49	<0.001
FMCE post-test (0–100)	1566	67.3	27	533	56.8	29	10.4	0.37	<0.001

The same trends exist for the shifts in IE 1 courses and IE 2 courses separately. Females always have more negative average shifts than males in both pedagogical approaches. There are some differences when comparing male and female shifts in IE 1 courses versus IE 2 courses. In the sense-making category, females have an average shift of −9% in IE 1 courses and an average shift of −14% in IE 2 courses (the difference is significant: $p<0.05$). In the conceptual category, males have an average shift of −3% in IE 1 courses and −6% in IE 2 courses, and in the applied conceptual category, males have an average shift of −2% in IE 1 courses and −7% in IE 2 courses (both differences are significant: $p<0.05$). For each of these differences, the shifts are more negative for IE 2 courses. For all other categories males and females have the same shifts in IE 1 and IE 2 courses. Aside from some small differences, partially and fully interactive courses have similar (negative) influences on students' attitudes and beliefs about physics.

B. Background differences

In Sec. III A we reported observed differences between males' and females' performance during the introductory physics course. Here, we examine the background and preparation of males and females. Male and female averages for each of the background variables collected are presented in Table IV. Note that not all data are available for all students, as is the case in any course. As a consequence of missing data the reported averages may be biased due to sampling error. We present them regardless as they are the best estimates we have of the values for all students who enroll in introductory physics.

Females have a higher average high school GPA than males by about 0.2 point. On average females take less high school physics than males, but they take about the same amount of high school calculus. These same data can be represented another way by looking at the percentage of males and females that have at least one year of high school physics and calculus. From Table V, 89% of males and only 80% of females in introductory physics completed at least one year of high school physics. Only small percentages, 16% of males and 11% of females, took two years of high school physics. There are only minor differences in the fraction of males and females who take high school calculus; 67% of males and 70% of females took at least one year of high school calculus. It is interesting to note that both males and females are more likely to take high school physics than to take high school calculus.

Males significantly outperform females on both the SAT-math and ACT-math tests. Surprisingly, there are no gender differences on either of the CU diagnostic exams that are given at the beginning of freshman year. Because of the differences in SAT and ACT scores, there is a significant difference between the average combined math scores of males and females.

IV. RESULTS: CORRELATION OF STUDENT BACKGROUND WITH STUDENT CONCEPTUAL PERFORMANCE

We have identified several aspects of the introductory physics course in which gender differences exist: conceptual surveys, course grades, attitudes and beliefs, and student background and preparation. The next step is to determine

TABLE V. Percentages of males and females that have taken high school courses in calculus and physics. The asterisk (*) indicates that the difference in percentages is significant via χ^2 test; $p<0.01$.

	% of males	% of females
1 year HS physics*	88.7	79.7
2 years HS physics*	15.5	10.7
1 year HS calculus	67.3	69.9

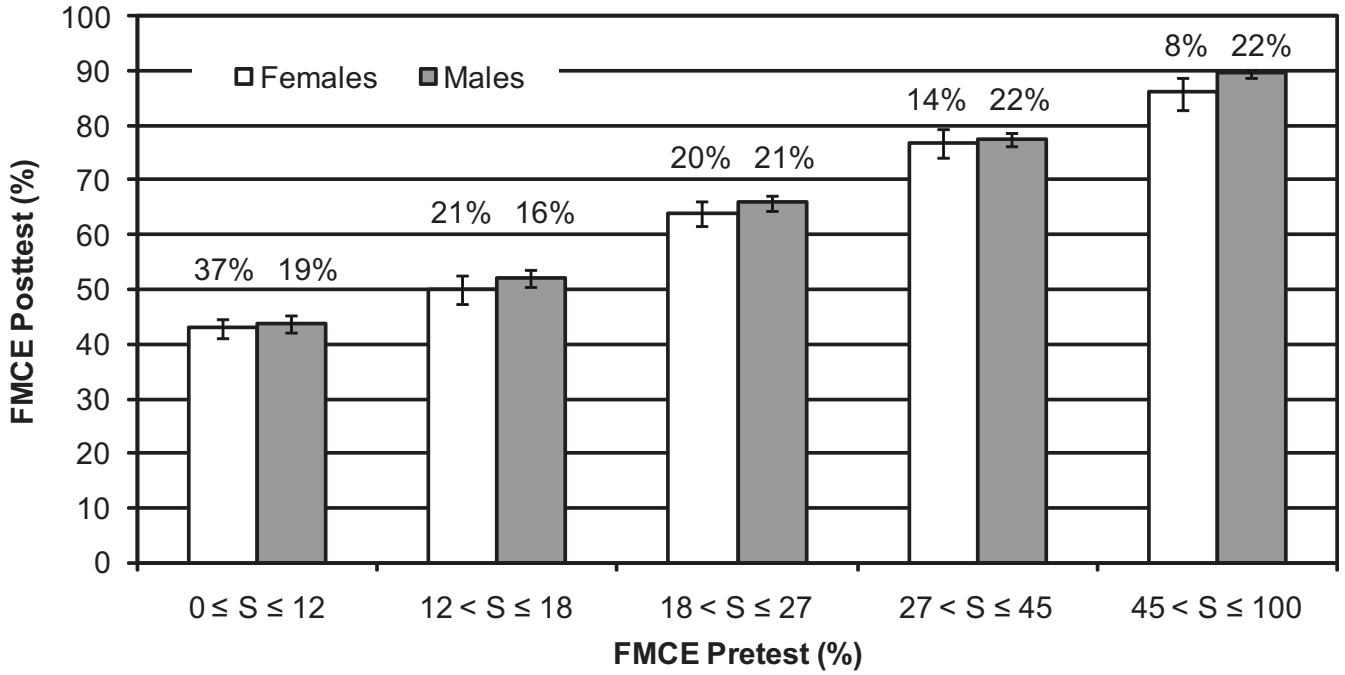

FIG. 4. Average FMCE post-test scores for females and males with matched FMCE pretest scores. The percentages above each bar represent the percentage of the females (or males) from the total in each bin. The error bars represent the standard error on the mean. There are no significant differences between males and females in any individual bin.

which, if any, of the student background factors are correlated with student performance on the conceptual survey and could therefore be contributing to the gender gap in observed post-test scores.

We first ask, do differences in male and female average post-test scores exist when students are grouped according to their pretest score? If males and females with similar pretest scores have different post-test scores, then there would be evidence that despite equal performance on measures of background physics knowledge, there is differential learning by gender. Students are binned by FMCE pretest score (each bin contains about equal numbers of students: $N \sim 420$), and then the average FMCE post-test score is calculated for males and females in each bin. The results are plotted in Fig. 4. The same trends that are described below exist for a range of reasonable bin sizes.

Students who have similar pretest scores have similar post-test scores regardless of gender. There were no statistically significant differences ($p > 0.1$) in any individual bin, i.e., between males and females who scored similarly on the pretest. Although the differences in each bin are not significant, males consistently score higher than females in all bins. We also see a correlation ($r = 0.56$) between FMCE pretest and post-test scores. These same trends exist for each individual semester.

We find that a higher percentage of the females fall into the low pretest bins. The percentages above each bar in the plot (Fig. 4) represent the percent of the females, or males, who fall into that bin. 58% of females versus 35% of males fall into the lowest two pretest bins, while 22% of females versus 44% of males fall into the highest two pretest bins. Thus, a dominant source of the observed gender difference (from Fig. 1) is attributable to the low pretest scores of females combined with the correlation between pretest and post-test scores.

The same trend exists for normalized learning gain and absolute learning gain; students with similar pretest scores have similar normalized and absolute gains regardless of gender. We also see a correlation, albeit weaker ($r = 0.3$), between FMCE pretest score and normalized gain and a correlation ($r = -0.2$) between FMCE pretest score and absolute gain. Furthermore, the results look the same whether students are in IE 1 or IE 2 courses.

To determine if taking high school physics influences the pre- and post-FMCE scores, we examine males and females who did and did not have a high school physics course. The results are presented in Tables VI and VII. Looking first at the pretest, students who had taken high school physics score significantly higher than students who did not have high school physics. The difference is greater for males than for females. It is also interesting to note that the gender gap for students who took high school physics is about 10 points, while the gender gap for students without any high school physics is only about 4 points. The gender gap on the post-

TABLE VI. Average FMCE pretest scores for males and females who did and did not take high school physics. The asterisk (*) indicates that the difference is statistically significant at $p < 0.01$.

	Males	Females	$M-F$
Had HS physics	33.5	23.9	9.6*
No HS physics	20.2	15.8	4.4*
Phys.–No phys.	13.3*	8.1*	

TABLE VII. Average FMCE post-test scores for males and females who did and did not take high school physics. The asterisk (*) indicates that the difference is statistically significant at $p<0.01$.

	Males	Females	$M-F$
Had HS physics	68.0	58.9	9.1*
No HS physics	60.7	44.9	15.8*
Phys.–No phys.	7.3*	14*	

test (Table VII) for those students who take high school physics is 9 points, statistically the same as the pretest gap. But, for students who had no high school physics, the gender gap on the post-test is 16 points (significantly larger than the pretest gap). Similar to the pretest, those students who took high school physics had higher average post-test scores than those students who did not. But, the gap is larger for females than for males.

V. RESULTS: ESTIMATION OF THE IMPACT OF STUDENT BACKGROUND ON THE GENDER GAP

A. Multiple regression analysis

Having identified several background variables that are correlated both with gender and with student performance on the FMCE, the next step is to model the post-test scores using multiple regression.[39] But due to the nature of our data, we cannot strictly interpret the statistical significance of the results, as they are likely to be biased. Because of ceiling effects, non-normal data, heteroskedasticity, and nonrandom sampling,[40] our data do not meet the strict assumptions of multiple regression that allow for unbiased interpretation of statistical significance. We can, however, use the regression analysis to describe the patterns in our data, without needing to meet the assumptions of multiple regression.[41] The results of the multiple regression analysis will describe the relationship between a student's post-test score and the values of several background variables for that student. Using this relationship, we estimate the difference in post-test scores for a male and female with all background variables being equal. In this way, we will determine how much of the gender gap can be accounted for by factors other than gender.

The post-test scores are modeled according to the equation

$$\mathrm{FMCEPOST} = b_0 + b_1 \times \mathrm{FEMALE} + \sum_{k=2}^{N} b_k \times \mathrm{VAR}_k,$$

where FMCEPOST is the post-test score on the FMCE, FEMALE is a dummy variable that is 1 for females and 0 for males, and VAR_k are the other background variables that are included in the model and any cross terms between FEMALE and the other background variables. b_k are the coefficients for each term, and the multiple regression analysis gives estimates for these coefficients. The coefficient of the FEMALE variable (b_1) gives the difference between a male's and a female's scores, with all other factors being equal. It is this coefficient that we are ultimately interested in.

We are modeling students' FMCE post-test scores rather than their absolute or normalized gain because we are primarily interested in reducing the gender gap in post-test scores. By modeling the post-test, we can determine what factors influence the post-test score and could therefore contribute to the gender gap. Each of the possible confounding variables is included in the regression analysis. Variables are entered sequentially in order to find the parsimonious combination of factors that best predicts the post-test score for each student. The best model will be judged based on the size of the coefficients, the increase in multiple R^2 (the fraction of variation in post-test scores that is accounted for by the variables in the model), and, to a lesser degree, the significance of variable coefficients (although, as mentioned above, the p values may be biased).

As stated above, not all data were available for all students. With this being the case, only a subsample of the students who took the introductory course was used in the multiple regression analysis. Recall that only 2099 of the 3728 students who enrolled in introductory physics between Spring 2004 and Spring 2007 took the FMCE pretest and post-test. Of these 2099 students, complete data[42] were available for 1027 students. These 1027 students make up the sample used in the analysis. It is important to keep in mind that the sample used in this analysis is not representative of all students who enroll in introductory physics. The percentage of females in this sample is 29%, which is higher than 24% for the population. It appears that females are more likely to take voluntary surveys (such as the FMCE and CLASS) which results in a slight oversampling of females. Also, the average course grades of females and males are higher for students in this sample than for students not in the sample. We again point out that by looking only at this sample of students we may be underestimating the gender gap. Furthermore, the results that we report below apply only to students in the sample and cannot be extrapolated to describe students not in the sample.

The results of the multiple regression analysis are shown in Table VIII. Four models are reported, starting with a bivariate model that includes only gender and then additional variables are added in each successive model. The table contains the coefficient estimates (b_k) for each model as well as the model-level statistics. The variables that are entered in each successive model are not only significant, but they also increase R^2 substantially (the additional variance explained by each model is significant via F test at the $p<0.01$ level). The R^2 for the final model is 0.44, such that the variation in the independent variables explains 44% of the variation in post-test scores.

We are interested in the difference between males' and females' post-test scores after controlling for several prior factors. In model 1, where only FEMALE is included as an independent variable, the gender difference is 10.7 points. This is just the average difference in post-test scores between males and females in this sample. In model 2, several covariates that are correlated with the post-test are added. When previous physics knowledge (FMCE pretest), previous math knowledge (combined math score), and previous attitudes and beliefs (CLASS pretest) are controlled, the gender dif-

TABLE VIII. Coefficient estimates and multiple regression model statistics for each multiple-regression model.

	Model 1	Model 2	Model 3	Model 4
Model-level statistics				
Multiple R^2	0.03	0.42	0.42	0.44
F statistic p value	<0.001	<0.001	<0.001	<0.001
Residual standard error	27.3	21.2	21.1	20.9
Predictors	b_k	b_k	b_k	b_k
Intercept	67.2	29.8	31.5	32.9
FEMALE	−10.7	−4.3	−10.0	−9.2
FMCE pretest		0.63	0.59	0.59
Combined math score		7.4	7.2	7.2
CLASS pretest		0.25	0.24	0.26
Semester B (IE 1)				1.3
Semester C (IE1)				−5.6
Semester D (IE 1)				−8.7
Semester E (IE 2)				−2.9
Semester F (IE 2)				−0.93
FEMALE × FMCE Pretest			0.23	0.20

ference drops to 4.3 points. Already, there is a substantial reduction in the gender difference once previous physics and math knowledge and attitudes and beliefs are accounted for.

When regressing post-test on pretest for males and females separately, we observe that the two regression lines have different slopes. Model 3 includes an interaction term that allows the slope of the FMCE pretest variable to differ for males and females. This term is the product of the two variables FEMALE and FMCEPRE. Since FEMALE is 0 for males and 1 for females, the interaction term is 0 for all males and is equal to the pretest score for all females. The inclusion of the interaction term in the model suggests that the pretest score differently predicts the post-test score for males and females. The interaction term also needs to be taken into account when estimating the gender difference. The gender difference now depends on the pretest score.

To get a final estimate of the gender difference, we turn to model 4. In this model, variables are added to take into account the semester that students took introductory physics. Controlling for semester is important for two reasons. First, by including a variable that controls for the semester that they took physics, some dependence among students due to taking physics at the same time is eliminated. Second, the average post-test scores are different in each semester. Including a semester variable will account for any differences that happen by semester which contribute to the post-test scores. Although we have no further information about specific aspects of each semester that could contribute to the differences, by including the semester variables we can see if there are differences once other prior factors are accounted for. The base case in model 4 is semester G (meaning there is no variable included for this semester). This means that the coefficients of each semester variable give the average difference between semester G and that semester after all other variables have been accounted for. For example, controlling for pretest, math knowledge, and attitudes and beliefs, the average difference between semester G and semester D is about −8.7 points. Note that some of the differences are substantial, suggesting that even though the courses look similar according to the curriculum, how the curriculum is enacted may differentially influence whether students learn.[35]

With model 4, a final estimate of the difference between a male's and a female's post-test scores, controlling for several other factors, can be estimated. This difference is given by

$$M - F = 9.2 - 0.2 \times \mathrm{FMCEPRE}.$$

The average pretest score for this sample is 30.3. The gender difference for a male and a female with the average pretest, and all other variables equal, is 3.2 points. This is a substantial reduction from the 10.7 point difference that is observed just by subtracting the average male and female post-test scores. Controlling for student background, we account for 70% of the observed gender gap. The effect size

$$\mathrm{ES} = \frac{\langle \text{post-test} \rangle_M - \langle \text{post-test} \rangle_F}{\mathrm{SD}_{\text{post-test}}}$$

went from 0.39, when no background variables were controlled, to 0.11, when measures of student background are controlled.

The resulting expression for the gender gap predicts that for males and females with pretest scores above about 45%, the gender gap reverses sign. Females with pretest scores greater than 45% are predicted to have *higher* post-test scores than males with the same pretest. While this result is encouraging, we need to be cautious. There are very few data for students with pretest scores above 45%, especially for females. Only 8% of the females and 20% of the males in the sample have pretest scores greater than 45%. Because there are not many data for students with higher pretests, we cannot be sure that the predictions made in this region are accurate. The same can be said for very low pretest scores. Only 10% of the students have pretests lower than 10%.

While the final model includes many variables that one might suspect would influence post-test scores, there are several variables that are not included. All of the variables listed above (Tables I and IV) were included in the analysis, but none were found to contribute significantly to the model beyond those variables already included in model 4. Looking more closely at some of the variables that are not included in the final model offers additional information.

Years of high school physics were somewhat correlated with the post-test ($r=0.2$) but were also correlated with the pretest ($r=0.3$). For this reason, we suspect that years of high school physics and the pretest were contributing some of the same information about the post-test score. Because the pretest was more highly correlated with the post-test ($r=0.6$), we chose to include that in the model over years of high school physics. In addition, others have pointed out that the specifics of the high school physics class are important,[6,18,19] and that information may have been more useful in the model than just years of high school physics. Similar conclusions were drawn about years of high school calculus and the

combined math score. There is not very much variation in high school GPA (mean=3.7, SD=0.3), as only students who were admitted to CU and who took an introductory physics course are included. We suspect that the lack of variation in GPA and its low correlation with post-test score ($r=0.1$) made it less likely to be a useful predictor of post-test score.

Students' declared major also was not a significant predictor of post-test score. This suggests that after accounting for background differences, there is no difference in the post-test scores of students by major. We also found that ethnicity was not a significant predictor. We suspect that this is largely due to the small numbers of minority students who enroll in the physics course at CU, which makes it difficult to accurately estimate the influence of ethnicity on post-test score.

There were also several interaction terms that we attempted to include in the model. Notably, we included an interaction between the variable FEMALE and each semester variable. None of these interactions was significant, meaning that the gender gap was the same in each semester after controlling for previous knowledge and attitudes. This suggests that differences in the post-test gender gap from semester to semester (Fig. 2) can be accounted for by differences in students' previous knowledge and attitudes. Interaction terms between gender and combined math score and between gender and CLASS score were also included, but neither contributed to the post-test score. This suggests that math score and CLASS score equally impact the post-test score for males and females.

B. Logistic regression analysis

The previous multiple regression analysis gave us only a description of the data. Another way to analyze the data that will allow for interpretation of the statistical significance of the results is to use logistic regression analysis.[43] It is used when the outcome variable of interest is a categorical variable rather than a continuous variable, for example, passing or failing rather than a raw score. While using this method allows us to make statistical claims, we lose the ability to predict students' actual post-test score and can only predict whether they will score above some threshold. What is gained in statistical specificity is lost in richness of the data analyzed. To model the data using logistic regression, the FMCE post-test variable is converted into a categorical variable (with any reasonable number of categories). The analysis was run for several threshold values (20%, 40%, 60%, and 80%) and also for several numbers of post-test categories (2–5). The results were similar for all threshold levels and number of categories, so we present the results for a threshold of 60% and two post-test categories here.

The frequencies of males and females who score above and below 60% on the FMCE post-test are presented in Table IX. Note that we are using the same sample of 1027 students that was used in the multiple regression analysis above. We observe that 64% of the males and 49% of the females score above 60% on the post-test. This difference in percentages is significant (via χ^2 test, $p<0.01$). Males and females are not equally likely to score above 60%. A gender gap is present in this measure of student physics knowledge at the end of the course. This difference is the gender gap that we are concerned with in the logistic regression analysis.

TABLE IX. Percentages of males and females who score above and below 60% on the FMCE post-test. These are the percentages for the sample of 1027 used in the logistic regression analysis. The difference in percentage of males and females that score above and below 60% is significant ($p<0.01$).

	Males (%)	Females (%)
FMCE post-test >60%	64.3	48.5
FMCE post-test <60%	35.7	51.5

In logistic regression, rather than modeling the raw dependent variable, the *logarithmic odds* of the dependent variable is modeled. In this context, odds is defined[43] as the probability of an event occurring divided by the probability of an event not occurring. The post-test data are modeled according to the equation

$$\ln[\text{odds}(\text{FMCEPOST} > 60\%)] = b_0 + b_1 \times \text{FEMALE} + \sum_{k=2}^{N} b_k \times \text{VAR}_k.$$

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. To determine the difference in odds, we are again interested in the coefficient of the FEMALE variable, b_1. The odds for a male and a female, all other variables being equal, are related according to the equation

$$\text{odds}_F = e^{b_1} \times \text{odds}_M.$$

The logistic regression analysis estimates the coefficients of each variable (as in the multiple regression analysis), which then allows a prediction of each student's odds of scoring above 60%. In each model we are interested in (1) whether the coefficient of FEMALE is significantly different from zero (as indicated by the p value) and (2) whether e^{b_1} is less than, greater than, or equal to 1. The results of the logistic regression analysis are shown in Table X. For each model the coefficient estimates (b_k) and p values are given, as well as an evaluation of e^{b_k}. Only gender is included in model 1. In this model, the coefficient of FEMALE is significantly different than zero ($p<0.001$), and we see that the odds of a female scoring above 60% are about half the odds of a male scoring above 60%. Just as in multiple regression, when only gender is included in the model, the predicted difference between males and females is just the observed difference.

In model 2, covariates of the post-test are controlled for, including prior physics and math knowledge, and prior attitudes and beliefs. In this model, the coefficient of FEMALE is not significant ($p>0.1$), meaning the odds of scoring above 60% for a female are not statistically different than the odds for a male, holding all other variables constant. A male

TABLE X. Logistic regression analysis results. In each model covariates are included to control for differences in student background. Gender is not significant in the final model.

	Model 1			Model 2			Model 3			Model 4		
Model-level statistics												
Pseudo-R^2 (Nagelkerke's)	0.03			0.43			0.43			0.45		
Likelihood ratio p value	<0.001			<0.001			<0.001			<0.001		
Predictors	b_k	**Sig.**	e^{b_k}	b_k	**Sig.**	e^{b_k}	b_k	**Sig.**	e^{b_k}	b_k	**Sig.**	e^{b_k}
Intercept	0.59	<0.001		−3.0	<0.001		−2.9	<0.001		−2.9	<0.001	
FEMALE	−0.65	<0.001	0.52	−0.26	0.13	0.77	−0.53	0.18	0.59	−0.23	0.19	0.80
FMCE pretest				0.08	<0.001	1.1	0.08	<0.001	1.1	0.08	<0.001	1.1
Combined math score				0.56	<0.001	1.7	0.55	<0.001	1.7	0.55	<0.001	1.7
CLASS pretest				0.02	<0.001	1.0	0.02	<0.001	1.0	0.02	<0.001	1.0
Semester B (IE 1)										0.14	0.59	1.1
Semester C (IE 1)										−0.68	0.01	0.51
Semester D (IE 1)										−0.75	0.01	0.47
Semester E (IE 2)										−0.10	0.71	0.91
Semester F (IE 2)										−0.06	0.81	0.94
FEMALE × FMCE Pretest							0.01	0.45	1.0			

and a female with the same background (as measured by the prior factors included in the model) are equally likely to score above 60% on the post-test.

The interaction term between the pretest and gender was included in model 3.[44] Unlike the results using multiple regression, here there is no significant interaction between prior knowledge and gender. The pretest has the same effect on whether students score above 60% on the post-test for both males and females. Because the interaction term is not significant, we do not include it in model 4. Again, the semester variables are included in model 4. Note that including the semester variables does not have a substantial impact on the coefficient of FEMALE; it remains insignificant, but it does allow us to compare odds of students across semesters. The base case is again semester G. Just as with the multiple regression analysis, there are some semesters (semesters C and D) in which the odds of scoring above 60% are significantly different from the odds in semester G. This regression analysis only allows for statistical comparison between semester G and the other semesters. Repeating the logistic regression analysis with all of the other semesters as base cases, we find that the odds in semesters C and D are statistically equal to one another but statistically different from the other four semesters. The odds in semesters B, E, F, and G are all statistically equal once prior factors are accounted for. Again, even though prior factors are accounted for, there are still statistically significant differences between some semesters.

With semester differences accounted for, the final estimate of the relationship between the odds for a female and the odds for a male, holding all other variables constant, is

$$\text{odds}_F = 0.8 \times \text{odds}_M.$$

This is smaller than 1 (but not statistically different from 1), meaning that the odds for males and females are statistically equal. By accounting for student background, the factor relating the odds of males and females has gone from 0.5 to 0.8. Using logistic regression and controlling for student background, we account for 60% of the observed gender gap in odds. The gender gap in odds can be largely accounted for by prior physics and math knowledge and prior attitudes and beliefs.

The variance explained by the final models is about 45% of the variance in post-test scores.[45] There may be other prior factors that we have overlooked that could be important in helping to explain the post-test score and could contribute to the gender gap. Notably, we have not included any variables that characterize students' motivations, study or learning habits, or their reasons for being in the class. Socioeconomic status is a demographic variable that was not included in the model. A proxy for socioeconomic status (financial aid information) was available but only for those students who applied for need-based financial aid, which was a too limited sample to include in the analysis. There are also other aspects of students' background that were not included (other high school courses, grades in high school courses, other components of standardized tests, etc.). As mentioned above, there are specific aspects about how a faculty member implements the curriculum that have not been accounted for. Only an overall semester variable was included, which does not contain more detailed information about how the curriculum was implemented.

There are limitations in the applicability of these regression results to the entire population of students in the study. The sample of students used in the regression analyses is only about 30% of the students that enrolled in introductory physics during the semesters included in the study. We reiterate that although the students in the sample are different from the population of all students, by using this sample of mostly high performing students (in terms of course grades)

and given the larger gender gap in course grades among the students not in the sample, it is possible that we are underestimating the gender gap of all students.

Finally, there are potential limitations due to the reliability of the instruments that were used to assess learning and prior knowledge. While our attempts to examine shifts in student performance allow us to normalize students against themselves, the broader scale concerns about gender-based biases, such as stereotype threat,[46] still remain. Some hint of test taking being a factor that differentially impacts female performance is the data on student grades. Consistently the males outperform the females on exams while females outperform males on homework and other course components.

VI. DISCUSSION AND CONCLUSIONS

While the differential performance of male and female students is now well documented, the sources of the gender gap and routes to addressing this disparity have been less well understood. By examining the performance and background of nearly 4000 students who took introductory physics at the University of Colorado, we begin to understand the sources of and possible solutions to this challenge. Our present studies find that the gender gap exists well beyond measures of student conceptual learning. Student grades vary by gender, both in overall scores and by course component. We observe that males and females have different shifts over the course of the semester in their attitudes and beliefs about physics, suggesting that males and females are experiencing the same course in different ways. The physics and mathematics background and preparation of students coming into our courses also vary by gender.

In taking a closer look at the gender gap in measures of conceptual performance, we observe that the pretest and post-test gender gaps are not consistent from semester to semester. Although the regression analysis suggests that these differences in the *gender gap* from semester to semester can be accounted for by background, differences in the average post-test score of all students from semester to semester are present even after controlling for student background. Given that there is relative consistency on the large scale in these courses, it appears that instructor differences, the course specifics, the way in which the curricula are implemented, and, potentially, the fine-grained choices that are made with regard to content and course structure impact the overall performance of all students. While we observe differences in males' and females' post-test scores and in their normalized gains, we find no significant differences in average absolute gain on these measures of conceptual learning in any semester or overall. If learning is defined by absolute gain, rather than normalized to prior knowledge, there is no gender gap.

Several of the background measures correlate with student performance on the FMCE post-test, suggesting that part of the gender gap may be attributed to differences in student background. In particular, when we bin students by pretest score, we find no difference in post-test scores between males and females with similar pretest scores. This is not the case when only taking into account whether or not students took high school physics. The gender gap in post-test scores is present both between males and females who did take high school physics and those who did not. Furthermore, the gender gap is exacerbated for those students who did not take high school physics. While controlling for whether students take high school physics does not account for the observed gender gap, our data suggest that differences in students' pretest scores and other measures of student background may account for a substantial fraction of the gender gap.

Both the multiple-regression and the multiple regression models confirm this interpretation, showing that a majority of the gender gap can be accounted for by factors other than gender explicitly. From the multiple regression analysis we find that only 3 points of the 11 point gender gap cannot be accounted for by background factors. From the logistic regression analysis we find that the odds of a male and a female scoring above 60% on the post-test are not statistically different once background factors are accounted for. Taken together, the results of these models suggest that the persistence of the gender gap is due in large part to differences in males' and females' preparation and background coming into the introductory course and not explicitly due to their gender.

In one sense, it may be interpreted that gender does not play a role in measures of student achievement—the variation in FMCE post-test score may be attributed to other variables, notably pretest score, student beliefs, and math achievement. For a given semester, male and female students make statistically indistinguishable grades. Such a stance would suggest that there is no explicit gender bias in the classes observed. Both males and females show learning gains from pretest to post-test. Nonetheless, in these classes we observe a gap in performance by gender and observe instances where, over the course of instruction, this gap is increased.

Another interpretation is that of implicit bias—that is, those components of a class that are most heavily weighted and essential for success disproportionately favor male students. While course grades are statistically neutral overall for a given semester, male students are more likely to score higher on exams (which are weighted more heavily in a typical class). Further, over all semesters we find a small but significant difference in overall course grades. While the classes studied are introductory courses with no expectation of prior knowledge of physics, those students who arrive to the class with greater background knowledge (higher pretest scores) are more likely to achieve high post-test scores and greater normalized learning gains. The class favors those students with stronger physics and math backgrounds—in this case, male students.

Such an arrangement of a class (or any social environment) plays to certain student backgrounds and when those backgrounds are correlated with particular demographic groups, it demonstrates bias. That is not to say this is an explicit or purposeful bias, but rather one that is the codified structure of systemic cultural bias.[47] Tatum[48] refers to this as a "smog of bias" and others to the privileged preparation of some group (at the expense of others) as an "accumulated disadvantage."[49] 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. Simply enacting research-based reforms, or supporting current practices (the *status quo*), may improve aggregate student learning gains but may also be promulgating the disparity of performance and lack of equity in our educational system.

ACKNOWLEDGMENTS

We gratefully acknowledge valuable assistance from the members of the Physics Education Research group at Colorado, the Discipline-Based Education Research group at Colorado, and the following physics faculty that were involved in our study: D. Anderson, D. Dessau, M. Holland, E. Kinney, S. Robertson, C. Rogers, J. Shepard, and J. Smith. Thanks to D. Briggs and J. Watkins for the help with statistical techniques and analyses. This work is supported by the APS, AAPT PhysTEC project, the National Science Foundation, and the University of Colorado. This material is based on work supported by the National Science Foundation under Grant No. REC 0448176, CAREER: Physics Education and Contexts of Student Learning.

[1] R. Ivie and K. N. Ray, AIP Report No. R-430.02, College Park, MD, 2005 (unpublished); www.aip.org/statistics

[2] R. R. Hake, in *Proceedings of the 2002 Physics Education Research Conference*, edited by S. Franklin, K. Cummings, and J. Marx (PERC Publishing, New York, 2002); http://www.physics.indiana.edu/~hake/PERC2002h-Hake.pdf

[3] M. Lorenzo, C. Crouch, and E. Mazur, Reducing the gender gap in the physics classroom, Am. J. Phys. **74**, 118 (2006).

[4] S. J. Pollock, N. D. Finkelstein, and L. E. Kost, Reducing the gender gap in the physics classroom: How sufficient is interactive engagement?, Phys. Rev. ST Phys. Educ. Res. **3**, 010107 (2007).

[5] E. Seymour and N. M. Hewitt, *Talking About Leaving: Why Undergraduates Leave the Sciences* (Westview, Boulder, CO, 1997).

[6] Z. Hazari, R. H. Tai, and P. M. Sadler, Gender differences in introductory university physics performance: The influence of high school physics preparation and affective factors, Sci. Educ. **91**, 847 (2007).

[7] K. K. Perkins, M. M. Gratny, W. K. Adams, N. D. Finkelstein, and C. E. Wieman, in *Proceedings of the 2005 Physics Education Research Conference*, AIP Conf. Proc. No. 818, edited by P. Heron, L. McCullough, and J. Marx (AIP, New York, 2006), p. 137.

[8] R. R. Hake, Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses, Am. J. Phys. **66**, 64 (1998).

[9] E. F. Redish, *Teaching Physics With the Physics Suite* (Wiley, Hoboken, NJ, 2003).

[10] L. C. McDermott and E. F. Redish, Resource Letter: PER-1: Physics Education Research, Am. J. Phys. **67**, 755 (1999).

[11] E. Mazur, *Peer Instruction: A Users Manual* (Prentice-Hall, Upper Saddle River, NJ, 1997).

[12] L. C. McDermott and P. S. Schaffer, *Tutorials in Introductory Physics* (Prentice-Hall, Upper Saddle River, NJ, 2002).

[13] P. Laws, P. Rosborough, and F. Poodry, Women's responses to an activity-based introductory physics program, Am. J. Phys. **67**, S32 (1999).

[14] M. Schneider, Encouragement of women physics majors at Grinnell college: A case study, Phys. Teach. **39**, 280 (2001).

[15] A. M. L. Cavallo, M. Rozman, and W. H. Potter, Gender differences in learning constructs, shifts in learning constructs, and their relationship to course achievement in a structured inquiry, yearlong college physics course for life science majors, Sch. Sci. Math. **104**, 288 (2004).

[16] The error bars in all figures represent the standard error on the mean. They are only a lower limit on the actual error. These error bars do not account for sources of error other than statistical error, such as systematic error or sampling bias.

[17] The fully interactive courses also have a higher average learning gain for both males and females than the partially interactive courses ($\langle g \rangle_{\text{partial}}=0.45$ and $\langle g \rangle_{\text{full}}=0.52$).

[18] P. M. Sadler and R. H. Tai, Success in introductory college physics: The role of high school preparation, Sci. Educ. **85**, 111 (2001).

[19] R. H. Tai and P. M. Sadler, Gender differences in introductory undergraduate physics performance: University physics versus college physics in the USA, Int. J. Sci. Educ. **23**, 1017 (2001).

[20] D. E. Meltzer, The relationship between mathematics preparation and conceptual learning gains in physics: A possible "hidden variable" in diagnostic pretest scores, Am. J. Phys. **70**, 1259 (2002).

[21] CAPA, http://www.lon-capa.org/; Mastering Physics, http://www.masteringphysics.com/

[22] V. Otero, N. D. Finkelstein, R. McCray, and S. Pollock, Who is responsible for preparing science teachers? Science **313**, 445 (2006).

[23] N. D. Finkelstein and S. J. Pollock, Replicating and understanding successful innovations: Implementing tutorials in introductory physics, Phys. Rev. ST Phys. Educ. Res. **1**, 010101 (2005).

[24] For instance, implementations of IE 2 curricula at Harvard University included cooperative problem-solving activities that were not part of IE 2 courses at CU.

[25] Preliminary analysis of the retention of physics majors suggests that only about 67% of those students who are declared physics majors in the first-semester course continue as physics majors in the second-semester course.

[26] R. K. Thornton and D. R. Sokoloff, Assessing student learning of Newton's laws: The force and motion conceptual evaluation and the evaluation of active learning laboratory and lecture curricula, Am. J. Phys. **66**, 338 (1998).

[27] Data on students' high school experience (high school courses, high school GPA, SAT-math, and ACT-math scores), student demographic data (gender, ethnicity, and declared major), and course data (semester of enrollment and course grade) were collected from the Office of Planning, Budget, and Analysis at CU–Boulder. None of these data were self-reported.

[28] For example, a student's SAT-math score S was converted to a z score using the transformation $z_S=(S-\langle S\rangle)/\mathrm{SD}$, where $\langle S\rangle$ is the average SAT-math score for all students and SD is the standard deviation of the SAT-math scores.

[29] W. K. Adams, K. K. Perkins, N. S. Podolefsky, M. Dubson, N. D. Finkelstein, and C. E. Wieman, New instrument for measuring student beliefs about physics and learning physics: The Colorado Learning Attitudes about Science Survey, Phys. Rev. ST Phys. Educ. Res. **2**, 010101 (2006).

[30] L. McCullough, Gender, context, and physics assessment, J. Int. Women's Stud. **5**, 20 (2004).

[31] D. Hestenes, M. Wells, and G. Swackhamer, Force Concept Inventory, Phys. Teach. **30**, 141 (1992).

[32] T. L. McCaskey, M. H. Dancy, and A. Elby, in *Proceedings of the 2003 Physics Education Research Conference*, AIP Conf. Proc. No. 720, edited by J. Marx, S. Franklin, and K. Cummings (AIP, New York, 2004), p. 37.

[33] E. Hazel, P. Logan, and P. Gallagher, Equitable assessment of students in physics: importance of gender and language background, Int. J. Sci. Educ. **19**, 381 (1997).

[34] Note that not all students who enroll in the course received a course grade. 163 students withdrew from the course (receiving a grade of W) and two students who took the course pass or fail (receiving a grade of P). Both students who took the course pass or fail took the FMCE. All but one of the students who withdrew from the course did not take the FMCE.

[35] C. Turpen and N. D. Finkelstein, in *Proceedings of the 2007 Physics Education Research Conference*, AIP Conf. Proc. No. 951, edited by L. Hsu, C. Henderson, and L. McCullough (AIP, New York, 2007), p. 951.

[36] A. F. Heckler (http://www.mps.ohio-state.edu/Personnel/Heckler/TauVsGpaper.final.pdf).

[37] The CLASS that was administered in semester A was the old version. Because of the many differences between the old and new versions, we excluded data from that semester from this analysis and future analyses where the CLASS was used.

[38] Shifts toward less expertlike beliefs are common, as documented in Ref. 29.

[39] R. G. Lomax, *An Introduction to Statistical Concepts for Education and Behavioral Sciences* (Lawrence Erlbaum, Mahwah, NJ, 2001).

[40] A nontrivial fraction of the students score high on the pretest, making their gain questionable (ceiling effect). The distribution of FMCE post-test scores is not Gaussian and is skewed toward high scores (non-normal data). The variance in the post-test scores is not constant across all pretest values (heteroskedasticity). The sample of students used in the multiple regression analysis was not randomly chosen (nonrandom sampling).

[41] R. A. Berk, *Regression Analysis: A Constructive Criticism* (SAGE, Thousand Oaks, CA, 2004), p. 212.

[42] Complete data means that data were available for at least one of the four math tests and all other variables in Tables I and IV. Also, semester A was excluded from both regression analyses; see Ref. 37.

[43] W. Mendenhall and T. Sincich, *A Second Course in Statistics: Regression Analysis* (Pearson Education, Upper Saddle River, NJ, 2003).

[44] The inclusion of the interaction term in model 3 changes the relationship between the odds for a female and the odds for a male. The relationship becomes $\mathrm{odds}_F=\exp(-0.534+0.013\times\mathrm{FMCEPRE})\times\mathrm{odds}_M$.

[45] An R^2 value of 0.45 is relatively high for regression analyses such as this. Similar analyses conducted by other researchers have R^2 values between 0.16 (Tai) and 0.36 (Sadler).

[46] S. J. Spencer, C. M. Steele, and D. M. Quinn, Stereotype threat and women's math performance, J. Exp. Soc. Psychol. **35**, 4 (1999).

[47] M. Dancy, in *Proceedings of the 2003 Physics Education Research Conference*, AIP Conf. Proc. No. 720, edited by J. Marx, S. Franklin, and K. Cummings (AIP, New York, 2004), p. 31.

[48] B. D. Tatum, *Why are All the Black Kids Sitting Together in the Cafeteria? And Other Conversations About Race* (Basic Books, New York, 1997).

[49] V. Valian, *Why So Slow: The Advancement of Women* (MIT, Cambridge, 1999).