Partitioning the gender gap in physics conceptual inventories: Force Concept Inventory, Force and Motion Conceptual Evaluation, and Conceptual Survey of Electricity and Magnetism

Rachel Henderson,[1,*] John Stewart,[2] and Adrienne Traxler[3]

[1]*Michigan State University, Department of Physics and Astronomy, East Lansing, Michigan 48824, USA*
[2]*West Virginia University, Department of Physics and Astronomy, Morgantown, West Virginia 26506, USA*
[3]*Wright State University, Department of Physics, Dayton, Ohio 45435, USA*

(Received 23 February 2019; published 28 May 2019)

Over the last decade, the "gender gap" in physics conceptual inventory scores has been extensively studied by the physics education research community. Researchers have identified many factors that influence the overall differences in post-test scores between men and women. More recently, it has been shown that the Force Concept Inventory (FCI) contains eight items that are substantially unfair; six are unfair to women, two are unfair to men. The Force and Motion Conceptual Evaluation (FMCE) and the Conceptual Survey of Electricity and Magnetism (CSEM), however, contain fewer unfair items. In this work, results from prior studies are used to further explore the gender gap in five large samples of conceptual inventory data: the FCI ($N_1 = 3663$), the FMCE ($N_2 = 2551$, $N_3 = 3719$), and the CSEM ($N_4 = 1767$, $N_5 = 2439$). The gender gap in these samples is partitioned into four components: the gender gap resulting from the student's academic performance, the gender gap resulting from prior preparation in physics, the gender gap resulting from instrumental fairness, and the gender gap of students with equal academic performance and physics preparation on the fair instrument. For all samples, very little of the gender gap was explained by differences in academic performance between men and women, measured by ACT or SAT math percentile scores or physics test average. The percentage of the gender gap resulting from instrumental fairness varied across samples from 30% in the FCI to 2% to 6% in the CSEM. A substantial part of the gender gap in four of the five samples (30%–40%) was explained by differences in prior physics preparation, measured by pretest scores on the conceptual inventories. Further correcting for conceptual physics prior preparation using the post-test score in the previous class reduced gender differences substantially.

DOI: 10.1103/PhysRevPhysEducRes.15.010131

I. INTRODUCTION

The "gender gap," gender differences between the scores of men and women on commonly used physics conceptual inventories, such as the Force Concept Inventory (FCI) [1], the Force and Motion Conceptual Inventory (FMCE) [2], and the Conceptual Survey of Electricity and Magnetism (CSEM) [3], has been thoroughly investigated. On average, men outperform women by 12% on the mechanics conceptual inventories and by 8.5% on electricity and magnetism conceptual inventories [4].

Many factors have been explored to explain the differences observed in the performance of men and women on conceptual physics evaluations. These factors may be broadly classified as factors related to general academic achievement, prior physics or mathematics preparation, and factors not related to achievement or preparation. Factors related to academic achievement include academic performance measured by course grades and tests of specific cognitive reasoning skills. A substantial body of research has demonstrated differences in academic course grades [5,6] with a consistent advantage to women. Extensive research has examined the differences between men and women on specific cognitive tasks [7–10] with women scoring generally higher on verbal reasoning tasks and men generally higher on spatial reasoning tasks. These differences can be very fine-grained with differences measured on related tasks in the same discipline [6]. Within physics, multiple academic, cognitive, and preparation measures have been used to explain gender differences including the Lawson test of scientific reasoning and the years of high school calculus as well as conceptual physics pretest score (see Madsen, McKagan, and Sayre, Table I for a summary [4]).

Factors not related to academic achievement and preparation have also been extensively examined; these include

*hende473@msu.edu

Published by the American Physical Society under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

TABLE I. Corrected pretest and post-test items. The items included on the corrected instruments.

Sample		Total	Items in corrected instrument
FCI-1	Pretest	10	1, 2, 3, 4, 7, 8, 10, 16, 19, 20
	Post-test	19	1, 2, 3, 4, 5, 7, 8, 10, 11, 13, 16, 17, 18, 19, 20, 25, 26, 28, 30
FMCE-2	Pretest	11	1, 16, 22, 23, 24, 26, 30, 31, 32, 34, 41
	Post-test	23	1, 2, 4, 14, 16, 17, 18, 19, 20, 22, 23, 24, 26, 30, 31, 32, 34, 36, 38, 40, 41, 42, 43
FMCE-3	Pretest	5	22, 24, 26, 31, 41
	Post-test	22	1, 2, 4, 14, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 30, 31, 32, 34, 38, 41, 43
CSEM-1	Pretest	12	1, 2, 4, 8, 9, 12, 13, 17, 18, 19, 30, 32
	Post-test	29	1, 2, 4, 5, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32
CSEM-3	Pretest	12	1, 2, 3, 5, 6, 8, 9, 12, 17, 18, 19, 30
	Post-test	31	All original items except 32

psychosocial factors and instructional factors. Psychosocial factors that have been shown to be related to gender differences in academic performance include science anxiety [11–13], mathematics anxiety [14,15], and stereotype threat [16]. Psychosocial factors have also been investigated as an explanation of performance differences in physics classes [17,18]. Classroom instructional mode and environment have also been explored as possible explanations of gender differences. The results of these studies have been inconsistent with some studies showing active-learning instruction produces decreased gender differences [19–21] while other studies show no effect of reformed instruction on gender differences [22–24].

For a more detailed discussion about the many sources that may influence the overall gender gaps on physics conceptual inventories, see Henderson *et al.* [25].

Performance differences between men and women on the individual items on physics conceptual inventories have been less thoroughly investigated. Much of the research in this area has focused on the FCI; Classical Test Theory [26–28] and Item Response Theory [27,29–31] have been used to examine the validity of the instrument. In addition, Differential Item Functioning (DIF) analysis has demonstrated that some of the items in the FCI are unfair to either men or women [30,32,33]. As previously described in Traxler *et al.* [33], "An item is defined as being "fair" if men and women of equal ability have the same chance of answering the item correctly." While less research has been performed on the FMCE and the CSEM, individual items on the FMCE [34–37] and the CSEM [38,39] have also been examined; however, most of this work was not differentiated by gender. Only one study has reported item-level gender fairness for the FMCE or the CSEM [40]. For a more complete discussion of item-level research on the FCI, FMCE, and CSEM, see Traxler *et al.* [33] and Henderson *et al.* [40].

In general, many factors have been shown to influence the overall gender gap, but physics education researchers have yet to come to an agreement as to the origin of these gender differences. This work presents an analysis which evaluates the relative importance of academic performance, instrumental fairness, and prior preparation on gender differences in the FCI, the FMCE, and the CSEM. It uses samples described in three previous studies [25,33,40] and attempts to shed additional light on the gender differences identified in these studies by modifying the conceptual instruments as proposed in those studies and by using relations between the student populations and instructional environments in the individual samples.

II. RESEARCH QUESTIONS

This study sought to the answer the following research questions:

RQ1: How much of the gender gap in physics conceptual post-test scores can be attributed to differences in general academic performance measured by ACT/SAT scores or physics test averages?

RQ2: How much of the gender gap can be attributed to instrumental fairness?

RQ3: How much of the gender gap can be attributed to differences in prior conceptual preparation in physics measured by pretest scores?

By answering these questions, this study forms a partition of the gender gap that may allow more targeted development of instructional interventions that will allow all students to succeed equally in physics classes. We end with some suggestions for instructors and researchers and a reminder that concept inventory gaps are only one element of the gender dynamics of a classroom.

We acknowledge that the model of a binary classification simplifies the complexity of gender identity [41]; however, this model is used throughout much of the physics education research (PER) literature that examines the gender differences found in the physics conceptual inventories. In addition, we were limited to the gender descriptions collected at the institutions studied. Future studies should explore the following results for other marginalized groups.

III. BACKGROUND

This section summarizes the results of three previous studies that examined the samples presented in this paper;

these works will be referenced as study 1, study 2, and study 3 in this work.

A. Study 1

In study 1, Henderson *et al.* examined the gender gap on the CSEM and found that men outperformed women by 5% on the pretest and 6% on the post-test [25]. This study also examined other qualitative and quantitative multiple-choice questions assigned in the course. A gender gap of 3% was also measured for qualitative lab quiz questions and qualitative test questions; however, men and women performed equally on the quantitative test questions. This result suggested that the gender gap in this sample could not be explained by psychological mechanisms such as science anxiety or stereotype threat. Why would a student experience stereotype threat on the qualitative questions on a test but not the quantitative questions on the same test?

Through a structural equation modeling analysis, a latent variable called conceptual physics performance/nonquantitative (CPP/NonQnt) was extracted. CPP/NonQnt represented the amount of conceptual performance that could not be explained by quantitative performance. The correlation between CPP/NonQnt and CSEM pretest score was larger for men ($r = 0.41$) than for women ($r = 0.20$) suggesting that the CSEM pretest was less predictive of CPP/NonQnt for women than for men. Study 1 presented a partial explanation of this effect by exploring the distribution of pretest scores. Women have 5% lower pretest scores on average than men; this produced a small shift in the distribution of pretest scores moving women slightly closer the binomial distribution of pure guessing scores. As such, it was much more difficult to distinguish moderately prepared women from unprepared women than it was to distinguish moderately prepared men from unprepared men.

The sample that was investigated in study 1 will be labeled "CSEM-1" in the current study. The number represents the institution from which the sample was collected.

B. Study 2

In study 2, Traxler *et al.* explored the validity and intrinsic bias of the FCI using Classical Test Theory and Item Response Theory [33]. The analysis identified many of the items on the FCI as problematic due to item difficulty and discrimination values outside the accepted range for well-functioning items. Study 2 also investigated item fairness employing both a graphical analysis and using DIF analysis. In the graphical analysis, five items stood out as significantly unfair to women: items 14, 21, 22, 23, and 27. DIF analysis showed that eight items were substantially unfair controlling for the student's overall post-test score, two of which were unfair to men.

To construct a fair, valid FCI, Study 2 iteratively removed unfair items until no items in the instrument showed bias. This process produced a 19-item instrument which, in turn, reduced the original gender gap by 50%.

Study 2 analyzed three samples from three different institutions. The largest sample from study 2 was also analyzed in the current study and is labeled "FCI-1."

C. Study 3

In study 3, Henderson *et al.* [40] replicated the fairness analysis of study 2 for the FMCE and the CSEM using two large FMCE samples and two large CSEM samples. Overall, there were fewer items in the FMCE and the CSEM that demonstrated substantial unfairness to either men or women. For the first FMCE sample in study 3, one item was substantially unfair to women, item 27_29. Study 3 used the modified scoring suggested by Thornton *et al.* [42] where some items were eliminated and some groups (clusters) were scored as a block. The notation 27_29 represents items 27, 28, and 29. In the second FMCE sample in study 3, two items were substantially unfair to women, item 27_29 and item 40.

For the CSEM, only one item, item 20 was substantially unfair and only for one of the two samples analyzed. This item was unfair to men.

Study 3 utilized four different samples all of which were further investigated in the current study. Sample 1 and sample 3A from study 3 are labeled "FMCE-2" and "FMCE-3," respectively, in the current study. Sample 2 and sample 3B are labeled "CSEM-1" and "CSEM-3," respectively.

In the current work, modified conceptual inventories were constructed which eliminated invalid or unfair items for all conceptual inventory pretests and post-tests for the samples used in these three studies. Hierarchical linear regression (HLR) was used to analyze the gender gaps controlling for academic performance, measured by test average or ACT/SAT math percentile, and prior physics preparation, measured by pretest scores. This allowed a "partitioning" of the gender gap to determine which factors were most important to the observed gender differences and whether the relative importance of the factors was consistent across instruments and institutions.

IV. METHODS

This work reports results for the FCI, the FMCE, and the CSEM; each of the analyzed samples were described previously in more detail in studies 1 to 3. Readers seeking more information about institution characteristics, sample characteristics, or instructional environment should consult these works.

A. Samples

This study utilized five samples collected at three different institutions. The institutions are denoted as University 1, University 2, and University 3. The samples

are denoted as FCI-1, FMCE-2, FMCE-3, CSEM-1, and CSEM-3 where the number represents the institution at which the sample was collected.

University 1: University 1 is a large southern land-grant university serving approximately 25 000 students. University level demographics for the undergraduate student population of University 1 were 76% White students, 4% African-American students, 9% Hispanic students, 4% students reporting two or more races, and other groups each with 3% or less [43].

University 2: University 2 is a large western land-grant university serving approximately 34 000 students. The demographic composition of the undergraduate population of University 2 consisted of 68% White students, 12% Hispanic students, 7% international students, 6% Asian students, 5% students reporting two or more races, and other groups each with 2% or less [43].

University 3: University 3 is a large eastern land-grant university serving approximately 30 000 students. The undergraduate demographic composition of University 3 consisted of 79% White students, 7% international students, 4% African-American students, 4% Hispanic students, 4% students reporting two or more races, and other groups each with 1% or less [43].

Sample FCI-1: Sample FCI-1 was collected at University 1. Data were collected in the introductory, calculus-based mechanics course, where the FCI was given as a pretest and post-test. Sample FCI-1 contains 3663 matched pretest and post-test pairs (77% men, 23% women). Sample FCI-1 is a subset of the sample investigated in study 2 where it was referenced as sample 1. The sample is smaller than that of the previous study because test average data were not available for all students. Students enrolled in this course participated in two 50-min lectures and two 2-h laboratory sessions each week. Throughout the period studied, the design of this course was stable; the course was overseen by the same instructor with attendance managed with a quiz. The laboratory sessions included multiple research-based techniques including small-group problem solving, hands-on inquiry-based explorations, and TA-led demonstrations.

Sample FMCE-2: Sample FMCE-2 was collected at University 2. FMCE pretest and post-test data were collected in the introductory, calculus-based mechanics course. Sample FMCE-2 contains 2551 matched pretest and post-test pairs (72% men, 28% women). Sample FMCE-2 is a subset of the sample analyzed previously in study 3 where it was referenced as Sample 1. The sample contains fewer records than that of the previous study because ACT/SAT scores were not available for all students. The course was presented with three 50-min lectures and one 50-min tutorial section each week. Four university faculty members taught the lecture sections using peer instruction with clickers. Within the tutorial sections, students worked the University of Washington *Tutorials in Introductory Physics* [44]. There was no laboratory associated with this course.

Sample FMCE-3: Sample FMCE-3 was collected at University 3. FMCE pretest and post-test data were collected in the introductory, calculus-based mechanics course. Sample FMCE-3 contains 3719 matched pretest and post-test pairs (79% men, 21% women). Sample FMCE-3 is identical to sample 3A in study 3. The instructional environment for sample FMCE-3 varied over the period studied. During all semesters studied, a learning assistant (LA) program [45] was implemented in the laboratory where research-based materials were presented in the laboratory. During the first half of the study, the course studied presented four 50-min lectures and one 2-h laboratory session each week with LAs provided to all labs. Many lecture instructors taught the class during this period. In the second half of the study, the class was revised to three 50-min lectures and one 3-h laboratory session each week with LAs provided to a subset of the laboratory sessions because of funding issues. The new structure was led by two co-instructors that implemented the same policies and employed peer instruction using clickers.

Sample CSEM-1: Sample CSEM-1 was collected at University 1. CSEM pretest and post-test data were collected in the introductory, calculus-based electricity and magnetism course. Sample CSEM-1 contains 1767 matched pretest/post-test pairs (77% men, 23% women). Sample CSEM-1 is a subset of the samples investigated in study 1 and study 3; in study 3 it was referenced as sample 2. The sample is smaller than that used in the previous studies because test average data were not available for all students. The instructional environment for sample CSEM-1 was similar to that of sample FCI-1. The course was led by one instructor and the instructional environment remained stable over the time period studied.

Sample CSEM-3: Sample CSEM-3 was collected at University 3. CSEM pretest and post-test data were collected in the introductory, calculus-based electricity and magnetism course. Sample CSEM-3 contains 2439 matched pretest and post-test pairs (81% men, 19% women). Sample CSEM-3 is identical to sample 3B in study 3. The instructional environment for sample CSEM-3 was similar to that of sample FMCE-3.

Many students matriculated from the mechanics course to the electricity and magnetism course at all the institutions studied. As such, the student populations of samples FCI-1 and CSEM-1 were similar as were the student populations of samples FCME-3 and CSEM-3.

B. Corrected conceptual inventories

For this analysis, the conceptual inventory pretest and post-test scores for each of the samples were modified. The modifications removed problematic items from the pretest and both problematic items and unfair items from the post-test as identified in studies 2 and 3. The scores after these

modifications are called "corrected" scores and the instruments, corrected instruments. To construct valid pretest scores for each instrument, items that were identified as problematic on the respective pretests for either men or women were eliminated. These items had difficulty or discrimination outside of the range suggested by Classical Test Theory. To correct the post-test scores, small to moderate and large DIF items identified by DIF analysis were removed, thus removing item-level unfairness from the instrument. Problematic post-test items as identified by Classical Test Theory were also removed. Study 2 did not find the same pattern of substantially unfair items in the FCI-1 pretest that were found in the post-test, as such, pretest scores were not corrected for fairness. Table I summarizes the included items on each of the valid pretests and the fair or valid post-tests.

C. Measures

Gender was coded dichotomously as the variable Gen with women coded as zero and men coded as one. General academic performance was represented by the variable APerf%. For FCI-1 and CSEM-1, APerf% was measured with the in-semester physics test average. The tests were approximately 70% quantitative and 30% qualitative and represented about 70% of the student's grade. For FMCE-2, FMCE-3, and CSEM-3, the ACT or SAT mathematics percentile score was used as the measure of academic performance. These percentile scores are represented by the variable ACTM% because the majority of the students took the ACT. When both scores were available, they were averaged. We acknowledge that physics test average and ACT/SAT mathematics score measure different facets of general academic achievement and that it would have been optimal if ACT/SAT mathematics scores had been available for all students. For a subset of Samples FMCE-3 and CSEM-3, both ACT/SAT scores and physics test averages were available allowing a comparison of the use of the two variables to measure general academic performance. While not identical, the partitions of the gender gap produced for students where both variables were available were very similar suggesting both variables measure academic performance similarly. This analysis is summarized Sec. V C and presented in detail in the Supplemental Material [46]. Pretest and post-test scores were converted to percentages and are represented by the variables Pre% and Post%.

All statistical analysis was performed in the "R" statistical software system [47].

V. RESULTS

Descriptive statistics for all samples are presented in Table II. Mean percentage score and standard deviation are reported for both the original, uncorrected instrument, and for the valid or fair corrected instrument.

A. Binning analysis

Many previous works investigating gender differences in conceptual inventory scores have employed binning, dividing students into subgroups with small ranges of pretest scores and calculating subgroup (bin) averages [20,21]. In all samples, there were pronounced differences in the distribution of men and women in the pretest bins. The percentage of women in a pretest bin decreased as the average score of the bin increased. A table of the distribution of men and women in each bin for the uncorrected instruments in presented in Table III; a similar table for the

TABLE II. Descriptive statistics. The mean and standard deviation of both the original and corrected instruments. All values are percentages presented as mean ± standard deviation.

	N	Uncorrected Pretest%	Corrected Pretest%	Uncorrected Post-test%	Corrected Post-test%	Test Average	ACT/SAT Math%
				FCI-1			
Men	2838	44.3 ± 18	53.4 ± 23	73.9 ± 17	73.2 ± 20	77.2 ± 13	⋯
Women	825	31.3 ± 14	40.3 ± 20	66.2 ± 17	68.7 ± 20	78.7 ± 13	⋯
				FMCE-2			
Men	1839	46.2 ± 28	50.3 ± 31	74.8 ± 26	74.9 ± 25	⋯	88.3 ± 11
Women	712	30.4 ± 22	35.4 ± 27	60.3 ± 27	63.0 ± 26	⋯	85.9 ± 13
				FMCE-3			
Men	2947	25.8 ± 20	41.3 ± 32	53.4 ± 28	53.5 ± 30	⋯	78.8 ± 16
Women	772	19.8 ± 14	33.1 ± 30	41.5 ± 24	42.6 ± 26	⋯	79.0 ± 16
				CSEM-1			
Men	1352	29.6 ± 11	39.2 ± 17	65.5 ± 16	65.2 ± 16	77.4 ± 13	⋯
Women	415	24.8 ± 8	35.4 ± 15	59.3 ± 16	59.0 ± 15	77.6 ± 13	⋯
				CSEM-3			
Men	1975	27.6 ± 11	44.0 ± 18	46.2 ± 18	46.6 ± 18	⋯	79.9 ± 15
Women	464	24.0 ± 9	37.7 ± 17	40.7 ± 17	40.6 ± 17	⋯	80.5 ± 15

TABLE III. Percentage of women by uncorrected pretest bin.

				FCI				
Bin	0–6	7–8	9–10	11–12	13–14	15–16	> 16	
FCI-1 (%)	50	33	26	23	14	14	6	
				FMCE				
Bin	0–4	5–6	7–8	9–10	11–12	13–14	15–16	> 16
FMCE-2 (%)	46	38	32	30	25	23	22	13
FMCE-3 (%)	26	22	20	19	11	14	10	9
				CSEM				
Bin	0–6	7–8	9–10	11–12	13–14	> 14		
CSEM-1 (%)	32	29	22	19	11	4		
CSEM-3 (%)	25	22	15	17	10	8		

corrected instruments is presented in the Supplemental Material [46].

Figures 1 and 2 plot the binned pretest scores against the post-test scores for all instruments; both the corrected and uncorrected plots are shown. Overall, except for some minor changes, the post-correction plots are very similar to the pre-correction plots. A linear regression line has been added for men and women to each plot; the regression was performed only including bins containing at least 30 students. Except for the FMCE-3 sample (which is problematic because of the very low number of retained items), the regression lines are striking in that they are nearly parallel. This suggests gender differences as a function of corrected pretest score could be investigated by simple linear models. The corrected CSEM-1 regression lines are more parallel than the uncorrected lines. The corrected FCI-1 regression line has a larger slope than the uncorrected line.

B. Partitioning the gender gap

This overall gender gap, δG, the difference in mean post-test score of men and women, observed in the uncorrected post-test scores could be produced by many factors. Hierarchical linear regression (HLR) analysis was used to determine the relative importance of each factor. The results of these regressions allow the partitioning of the overall gender gap δG in the uncorrected instrument into

- δG_{pop}, the gap resulting from differences in general academic performance between men and women measured by either ACT/SAT mathematics percentile score or test average, the population gap;
- δG_{fair}, the amount of the gap explained by correcting the instrument for fairness, the fairness gap;
- δG_{prep}, the part of the gap resulting from differences in physics conceptual preparation using the corrected pretest to measure preparation, the preparation gap; and
- δG_{equal}, the gap of men and women with equal academic performance and equal physics conceptual preparation on the valid or fair corrected instrument.

This combined model can be written as

$$\delta G = \delta G_{\text{pop}} + \delta G_{\text{fair}} + \delta G_{\text{prep}} + \delta G_{\text{equal}}. \quad (1)$$

The terms in Eq. (1) were calculated through two HLRs, one using the uncorrected post-test score as the dependent variable, the other using the corrected post-test score. The δG parameters are related to the regression coefficient of a dichotomous variable (Gen) coded as 0 for women and 1 for men. The dependent variable in the uncorrected regressions is the uncorrected post-test percentage, Post %, and the uncorrected pretest percentage is used as an independent variable, Pre%. The dependent variable in the corrected regressions is the corrected post-test percentage, Post^C%, with the corrected pretest percentage, Pre^C% as an independent variable. The superscript "C" was used to indicate corrected pretest percentage, post-test percentage, and regression coefficients calculated with these quantities. Three regressions were carried out for each dependent variable; the uncorrected regression equations are given by

$$\text{Post\%} = \beta_{11} + \beta_{12}\text{Gen} \quad (2a)$$

$$\text{Post\%} = \beta_{21} + \beta_{22}\text{Gen} + \beta_{23}\text{APerf\%} \quad (2b)$$

$$\text{Post\%} = \beta_{31} + \beta_{32}\text{Gen} + \beta_{33}\text{APerf\%} + \beta_{34}\text{Pre\%}. \quad (2c)$$

The variable APerf% measures general academic performance (ACT/SAT math score or physics test average), Pre % is the pretest percentage score, and Post% the post-test percentage score. The regression coefficients are β_{ij}, where i represents the model and j the term in the model. A similar set of regressions was carried out for the corrected pretest Pre^C% and post-test Post^C% with regression coefficients denoted by β^C_{ij}.

The variables used to partition the gender gap are summarized in Table IV. Table V presents the uncorrected regressions for Sample FCI-1. Each successive model in Table V adds another independent variable. The first model is labeled model FCI-1-1 indicating model 1 of sample

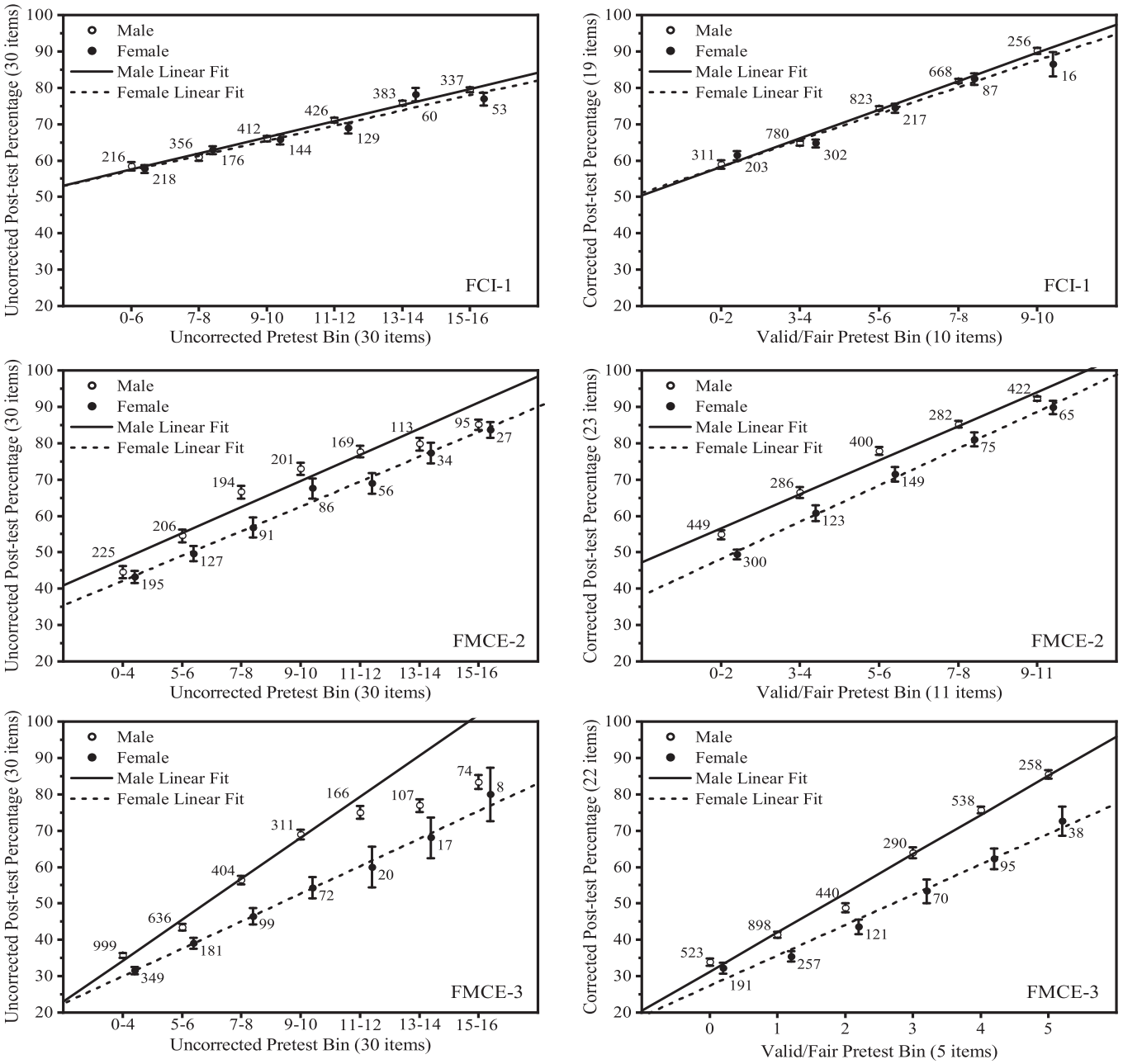

FIG. 1. Post-test vs pretest. Average pretest and post-test percentage scores on the uncorrected (left column) and corrected (right column) mechanics instruments. Linear fits only included bins containing greater than 30 students.

FCI-1. The overall gender gap δG is the regression coefficient β_{12} of model 1 with only Gen as the independent variable [Eq. (2a); model 1 for sample FCI-1]. The gender gap controlling for the academic performance of the students δG_A is extracted from Eq. (2b) (model FCI-1-2) which adds APerf% (test average in this sample) as an independent variable. The gender gap controlling for both prior physics preparation and academic performance $\delta G_{A,\text{prep}}$ [Eq. (2c); model FCI-1-3] adds the pretest percentage Pre% to model FCI-1-2. The corrected version of model FCI-1-1 is labeled FCI-1-1C. The corrected gender gap, δG^C, using the corrected instrument is calculated as the regression coefficient of Gen in model FCI-1-1C using the corrected post-test percentage Post^C% as the dependent variable. The corrected gender gap controlling for academic performance, δG_A^C, is the Gen regression coefficient in model FCI-1-2C. The corrected gender gap controlling for academic performance and prior physics preparation, $\delta G_{A,\text{prep}}^C$, is the Gen regression coefficient in model FCI-1-3C using Gen, APerf%, and corrected pretest percentage Pre^C% as independent variables.

The coefficients were then used to decompose the overall gender gap δG. The part of the overall gap that can be attributed to differences in the academic performance of the

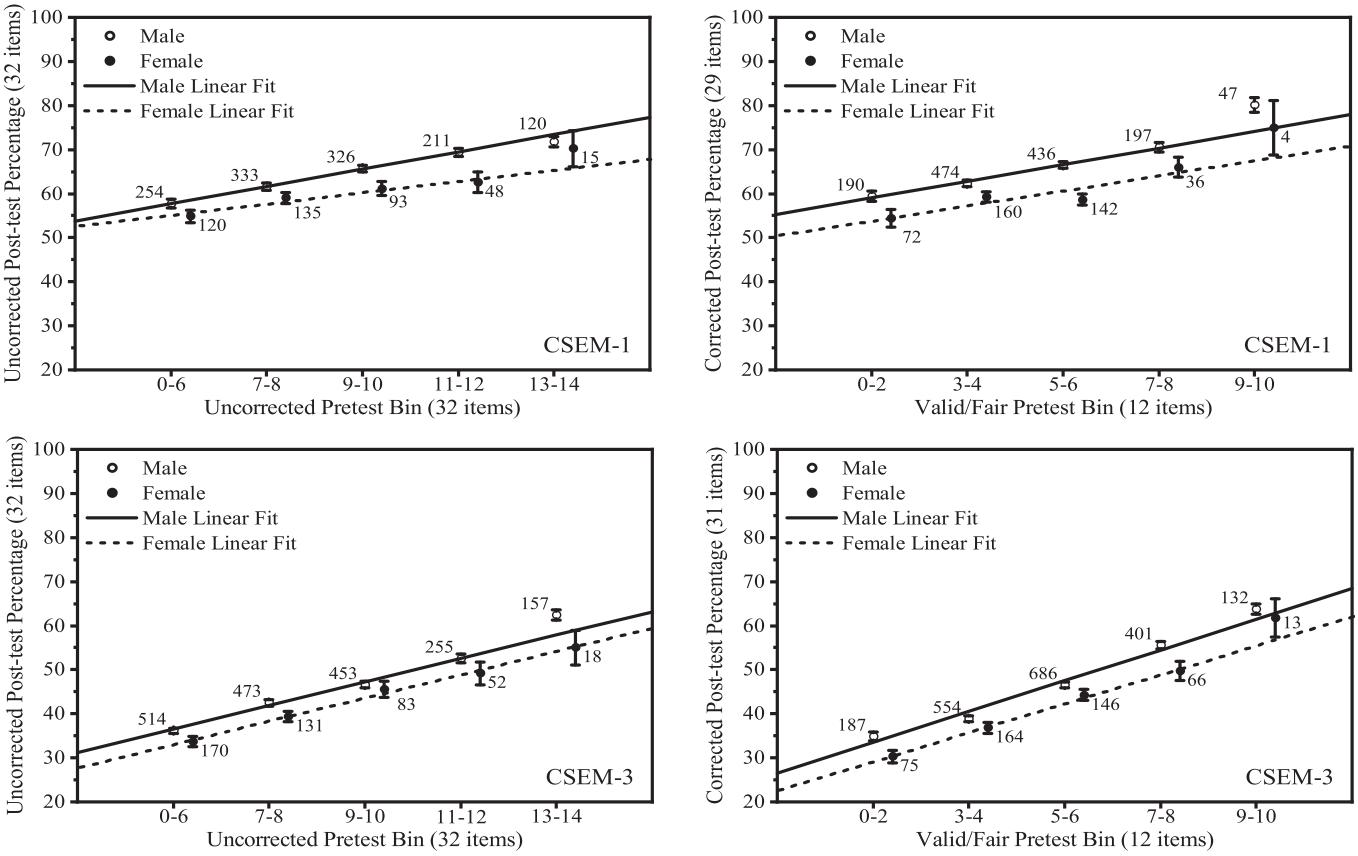

FIG. 2. Post-test vs pretest. Average CSEM pretest and post-test percentage scores on the uncorrected (left column) and corrected (right column) instrument. Linear fits only included bins containing greater than 30 students.

men and women in the samples (the population difference) is $\delta G_{\text{pop}} = \delta G - \delta G_A$. The amount of the gender gap attributable to the overall fairness and validity of the instrument is $\delta G_{\text{fair}} = \delta G_A - \delta G_A^C$, comparing the corrected and uncorrected instrument for students of the same academic performance. The amount of the gender gap attributable to physics preparation differences of students with the same academic performance on the corrected instruments (the preparation gap) is $\delta G_{\text{prep}} = \delta G_A^C - \delta G_{A,\text{prep}}^C$. The remaining gap (the fair, equally prepared and performing gap) $\delta G_{A,\text{prep}}^C = \delta G_{\text{equal}}$ is the gender difference attributable to equally prepared students with the same academic performance on the corrected instrument.

Table VI reports these quantities for all samples. Both standardized values (subtracting the mean and dividing by

TABLE IV. Summary of variables used in partitioning the gender gap.

Variable	Equation	Description
δG	β_{12}	Overall gender gap in the uncorrected post-test percentage.
δG_A	β_{22}	Gender gap correcting for academic performance on the uncorrected post-test.
$\delta G_{A,\text{prep}}$	β_{32}	Gender gap correcting for academic performance and physics preparation on the uncorrected post-test.
δG^C	β_{12}^C	Overall gender gap in the corrected post-test percentage.
δG_A^C	β_{22}^C	Gender gap correcting for academic performance on the corrected post-test.
$\delta G_{A,\text{prep}}^C$	β_{32}^C	Gender gap correcting for academic performance and physics preparation on the corrected post-test.
δG_{pop}	$\delta G - \delta G_A$	The part of the gender gap resulting from differences in academic performance.
δG_{fair}	$\delta G_A - \delta G_A^C$	The part of the gender gap resulting from the fairness of the instrument.
δG_{prep}	$\delta G_A^C - \delta G_{A,\text{prep}}^C$	The part of the gender gap resulting from differences in physics preparation.
δG_{equal}	$\delta G_{A,\text{prep}}^C$	The gender gap of equally prepared and performing students on the corrected instrument.

TABLE V. Hierarchical linear regression for the FCI (sample FCI-1). The superscript c denotes $p < 0.001$.

Model	Variable	B	SE	β	R^2	ΔR^2
		Uncorrected				
FCI-1-1	GenderM	7.76^c	0.68	0.44	0.034^c	
FCI-1-2	GenderM	9.07^c	0.53	0.52	0.428^c	0.394^c
	Test Ave.	0.84^c	0.02	0.63		
FCI-1-3	GenderM	3.98^c	0.50	0.23	0.540^c	0.112^c
	Test Ave.	0.65^c	0.02	0.49		
	Pretest%	0.37^c	0.01	0.38		
		Corrected				
FCI-1-1C	GenderM	4.48^c	0.78	0.23	0.009^c	
FCI-1-2C	GenderM	5.96^c	0.60	0.30	0.414^c	0.405^c
	Test Ave.	0.96^c	0.02	0.64		
FCI-1-3C	GenderM	2.65^c	0.59	0.13	0.475^c	0.061^c
	Test Ave.	0.81^c	0.02	0.54		
	Pretest%	0.24^c	0.01	0.27		

the standard deviation) using the standardized β coefficients and unstandardized values using the B coefficients are presented. The set of regressions used to calculate δG for the FCI in sample FCI-1 is shown in Table V; regressions for the other samples are presented in the Supplemental Material [46]. The table presents the regression coefficient B and its standard error SE, the standardized regression coefficient β, the variance explained by the model R^2, and the additional variance explained by a nested model ΔR^2.

Figure 3 presents a visual representation of the partitioning of the gender gap shown in Table VI. To create this representation, first the sum of the absolute value of each δG forming the partition was calculated to form the total absolute gender gap $|\delta G^T|$. The percentage of each partition was then calculated; for example, the percentage of the population gap was calculated as $100\%|\delta G_{\text{pop}}|/|\delta G^T|$. This somewhat circuitous calculation was needed to account for the negative gender gaps.

C. Comparison of academic performance measures

This study used both test average and ACT/SAT math percentile score as measures of academic performance. For a subset of the FMCE-3 and CSEM-3 samples, both variables were available allowing a comparison of the differences between these measures. The subsets contained 963 men and 271 women for FMCE-3 and 654 men and 171 women for CSEM-3. While both measures did not produce identical results, the resulting partition of the gender gap was very similar. As such, comparisons of the partition using different academic performance measures shown in Table VI should be valid. The detailed comparison is presented in the Supplemental Material [46].

D. δG_{equal}

The gender difference for equally prepared students with equal general academic performance, δG_{equal}, could depend on many factors; psychosocial factors and features of the instructional environment have been advanced to explain gender differences not related to academic performance or preparation. Psychosocial explanations of academic gender differences include stereotype threat, science anxiety, and math anxiety. Instructional factors include whether the courses used research-based practices. Both factors are reviewed in the introduction and more thoroughly in study 1. The causes of δG_{equal} almost certainly vary by student population and university environment; however, additional data and analysis provided in study 1 for the CSEM-1 sample make it difficult to support psychosocial factors as the cause of δG_{equal} for this sample. Study 1 also reported results for both quantitative and qualitative multiple-choice items that were not part of the CSEM including

TABLE VI. Decomposition of the gender gap. Presents a partitioning of the gender gap. All unstandardized values are percentages.

	Uncorrected			Corrected			Calculated			
	δG	δG_A	$\delta G_{\text{A,prep}}$	δG^C	δG^C_A	$\delta G^C_{\text{A,prep}}$	δG_{pop}	δG_{fair}	δG_{prep}	δG_{equal}
				Unstandardized						
Sample FCI-1	7.76	9.07	3.98	4.48	5.96	2.65	−1.31	3.11	3.31	2.65
Sample FMCE-2	14.49	12.80	5.27	11.94	10.35	4.60	1.69	2.45	5.75	4.60
Sample FMCE-3	11.99	12.10	7.30	10.84	10.96	7.47	−0.11	1.14	3.49	7.47
Sample CSEM-1	6.14	6.28	4.74	6.25	6.39	5.92	−0.14	−0.11	0.47	5.92
Sample CSEM-3	5.51	5.77	3.08	5.91	6.17	3.67	−0.26	−0.40	2.50	3.67
				Standardized						
Sample FCI-1	0.44	0.52	0.23	0.23	0.30	0.13	−0.07	0.21	0.17	0.13
Sample FMCE-2	0.54	0.47	0.20	0.46	0.40	0.18	0.06	0.08	0.22	0.18
Sample FMCE-3	0.43	0.43	0.26	0.37	0.37	0.25	−0.00	0.06	0.12	0.25
Sample CSEM-1	0.39	0.40	0.30	0.39	0.40	0.37	−0.01	0.00	0.03	0.37
Sample CSEM-3	0.31	0.32	0.17	0.33	0.34	0.20	−0.01	−0.02	0.14	0.20

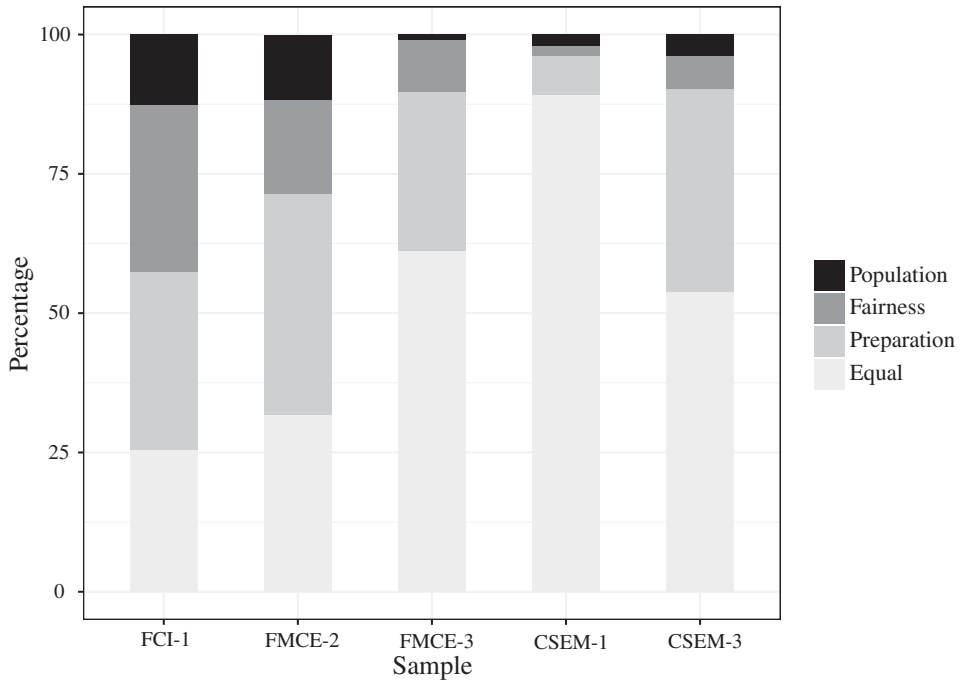

FIG. 3. Relative percentage of each partition.

quizzes given in the laboratory (lab quizzes) and qualitative and quantitative multiple-choice test questions. While a 3% gender difference with an advantage toward men was found in qualitative lab quizzes and qualitative test questions, no gender differences were found in quantitative test questions. The CSEM was given and graded as a lab quiz and there was a 6% gender difference on the post-test in this sample. The course instructor reported that both the qualitative and quantitative test items required a mix of verbal, logical, and graphical reasoning for their solution. The failure to observe a gender difference in the quantitative test items while observing gender differences in the qualitative test questions strongly suggests that psychosocial factors do not explain the gender differences. It is very hard to see how stereotype threat, for example, would function for qualitative items but not quantitative items on the same test. This suggests, for Sample CSEM-1, that δG_{equal} should be zero.

If $\delta G_{\text{equal}} = 0$, then we must revisit our assumption that academic performance, test fairness, and prior preparation have been correctly controlled. The DIF analysis used to produce the fair instruments is the standard method of ensuring fairness. It also seems very likely that the physics test average is an accurate measure of academic performance for this sample. As such, the assumption that the CSEM pretest score accurately measured prior preparation in physics must be reexamined. There is some support for challenging this assumption; study 1 showed that female pretest scores were much more weakly correlated with a latent variable measuring the student's qualitative performance not explained by his or her quantitative performance than male pretest scores.

Many theoretical objections can be raised for the assumption that CSEM (or FCI and FMCE) pretest scores are an accurate measure of prior preparation. The CSEM measures a limited subset of concepts in electricity and magnetism; this limited coverage may generate inaccurate results. The CSEM has very limited coverage of Newtonian mechanics and energy; these concepts are often used in conceptual electricity and magnetism problems. As such, the CSEM may not measure the student's mechanics preparation accurately. Further, and possibly most importantly, the CSEM pretest estimates the state of student knowledge early in the class and therefore only measures prior preparation that is directly retained. (Sec. VII has additional comments on concept inventories as measures of student knowledge or learning.) A pretest cannot measure the well-documented advantage to the student of relearning material rather than learning it for the first time [48–50].

To explore whether δG_{equal} was the result of prior preparation not captured by pretest scores additional measures of prior preparation were needed. For samples FCI-1, CSEM-1, FMCE-3, and CSEM-3, a subset of students completed both the mechanics and electricity and magnetism classes. For these students, either FCI or FMCE post-test results were also available, as well as CSEM scores. For sample CSEM-1, there were 1073 students for which a FCI post-test score was available (826 men, 247 women). Reproducing the partitioning of the gender gap shown in Table VI for this restricted sample yielded $\delta G = 4.90$, $\delta G_{\text{pop}} = -0.85$, $\delta G_{\text{fair}} = -1.60$, $\delta G_{\text{prep}} = 0.41$, and $\delta G_{\text{equal}} = 6.94$. If the FCI post-test score is used to measure prior preparation along with the CSEM pretest score [adding it as an independent variable to

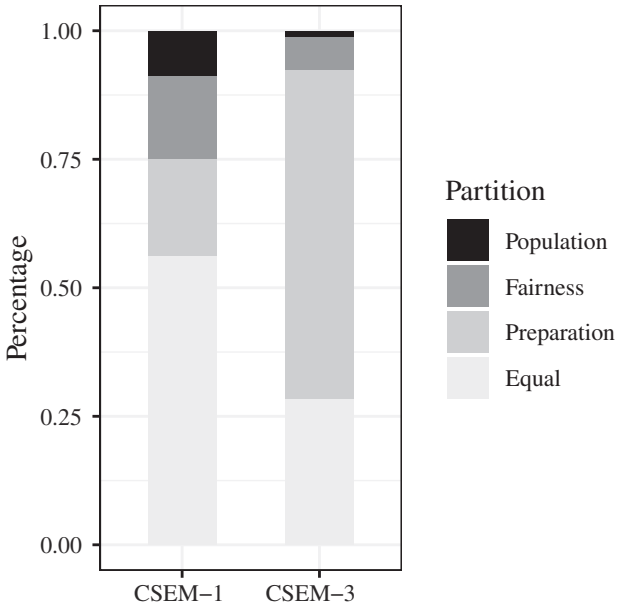

FIG. 4. Relative percentage of each partition using the mechanics post-test from the previous class as an additional measure of preparation.

Eq. (2c)], the last two terms change to $\delta G_{\text{prep}} = 1.84$ and $\delta G_{\text{equal}} = 5.51$. In this, adding FCI post-test scores as an additional measure of preparation reduced the equal gap by 1.43 or 21%. Figure 4 updates the CSEM results in Fig. 3 using the mechanics post-test results as an additional measure of prior preparation.

For the CSEM-3 sample, there were 1788 students for which a FMCE post-test score was also available (1413 men, 375 women). Reproducing the partitioning of the gender gap for this restricted sample yielded $\delta G = 6.01$, $\delta G_{\text{pop}} = 0.08$, $\delta G_{\text{fair}} = -0.44$, $\delta G_{\text{prep}} = 2.31$, and $\delta G_{\text{equal}} = 4.06$. If the FMCE post-test score is used to measure prior preparation along with the CSEM pretest score, the last two terms change to $\delta G_{\text{prep}} = 4.42$ and $\delta G_{\text{equal}} = 1.95$. Adding FMCE post-test scores as an additional measure of preparation reduced the equal gap by 2.11, or 52%.

VI. DISCUSSION

The primary results of this paper are captured in Fig. 3 and Table VI. For all samples, very little of the measured gender differences could be attributed to academic performance differences between men and women. Correcting the instruments for fairness explained differing amounts of the gender differences; fairness accounted for 30% of the gender difference in the FCI-1 sample, smaller but significant amounts of the gender differences in the FMCE (17% in FMCE-2 and 9% in FMCE-3), and little of the gender differences in the CSEM (2%–6%). This was fairly consistent with the size of the fairness effects calculated by the DIF analysis in studies 2 and 3. Prior preparation measured by pretest score explained consistently 30%–40% of the gender differences in all samples except sample CSEM-1. This left from 26% to 90% of the gender difference (δG_{equal}) unexplained. These percentages were calculated using the same method as was used to construct Fig. 3, by summing the absolute values of δG_i.

The 30% fairness gap measured in sample FCI-1 is smaller than the 50% reduction in the gender gap for the fair instrument presented in study 2. This difference results from the correction of the gender gap for academic performance which was performed in the current study but not study 2. The women in sample FCI-1 were higher performing than the men; correcting for this increased the gap.

There was a substantial difference in δG_{prep} (the gender difference resulting from prior preparation) between samples CSEM-1 and CSEM-3. This may partially be the result of the different prerequisite requirements of the classes. For CSEM-1, Calculus 2 is a co-requisite, while for CSEM-3, Calculus 2 is a prerequisite. As such, students in CSEM-3 are generally later in their academic career than students in CSEM-1. This suggests that some of the prior preparation difference may result from the student's experiences in college classes other than physics.

Analysis of a matched sample where both mechanics and electricity and magnetism pretest and post-test scores were available provided evidence that a substantial part of δG_{equal} could be explained by additional prior preparation measures. Study 1 further suggests that, to the extent that δG_{equal} results from psychosocial factors, that it should be zero for the CSEM-1 sample. This suggests that pretest scores, or at least CSEM pretest scores, do not provide an accurate measure of prior preparation. It is possible, as new measures are developed, that it will be shown that a substantial part of δG_{equal} is also the result of prior preparation.

FCI-1 and CSEM-1 have similar student populations and instructional environments. The large difference in δG_{equal} between the samples provides further evidence that δG_{equal} in the CSEM-1 sample cannot be explained by psychosocial effects. The instructional similarities between the two samples also suggest that δG_{equal} is not the result of instructional differences.

VII. IMPLICATIONS

This paper is one of a series that examined item-level gender fairness in several introductory physics conceptual inventories [25,33,40]. One of the original motivating questions for this research was how much of a widely discussed gender gap on these tests results from psychometric problems in the tests themselves? This question had been probed for the FCI ([33], Sec. I), but otherwise was largely open. In the course of investigating this larger issue, we have tried several methods to tease apart the sources of this gap: reduced or valid subsets of instruments, linear

modeling incorporating pretest scores and other preparation measures, and the framework in this paper of population, fairness, preparation, and equal partitions. We found that sometimes a small number of anomalous and problematic items could be identified [33], but often prior preparation was a larger contributor, and in some samples no combination of these factors can explain even half the gap.

Physics faculty use concept inventories for a number of reasons. Instructors may seek to gauge the quality of their teaching by comparing pre- and post-test scores, or may want to know how well students have learned the major concepts of the course. Researchers may want a standardized measure of the effectiveness of new curricula or interventions. Departments may suggest or require that instructors collect the data for official accountability or accreditation plans, or may simply value the practice of measuring teaching effectiveness. In all cases, it is important to remember that instruments must normally be calibrated before their readings make sense. This is as true for conceptual inventories as it is for probes on an oscilloscope.

It is also important to recall that multiple-choice tests, even those grounded in research on student thinking, are only one possible way to assess learning. They use questions to probe constructs such as "conceptual understanding of Newton's laws," which are defined with varying levels of theoretical clarity [51] and which may or may not emerge as expected by the test designers [52]. By the nature of their format, the information they provide is mediated by students' skill with taking standardized multiple-choice tests. Other researchers might prioritize different physics concepts or different ways of operationalizing them in test constructs [53], and instructors may value skills along other dimensions, such as constructing graphs or problem solving. These limits notwithstanding, conceptual inventories are used in many classrooms as the basis for claims about student learning. It is thus important to understand the complex range of factors that contribute to a single numerical score.

For instructors, the inventories we have examined—the FCI, FMCE, and CSEM—may completely encompass the content they want to assess. If they do not, other options exist that may be more appropriate (Madsen *et al.* [54] give a recent list). If instructors are giving points based on the number of correct items, they should check first for invalid items for their population. Though we found several consistently problematic items on the FCI, we also saw variation in the corrected instruments across our samples (Table I). The best way for instructors to understand the fairness of an instrument for their local population of students is to check the data. Calculating item difficulty using Classical Test Theory is straightforward, does not have the large sample size requirements of IRT models, and can still flag problematic items on pretest or post-test. The Classical Test Theory difficulty is simply the mean item score, allowing the graphical fairness analysis presented in studies 2 and 3 to be performed by any instructor with relatively small sample sizes.

For researchers, calibration is also essential if conceptual inventory scores are being used to make or bolster claims about effective materials. The process of calculating item difficulty, reducing instruments to valid items, and comparing new and old percentage scores (Table II) is one way to check whether problematic items are influencing claims about student learning. Researchers may also be one of the drivers of departmental data collection, and may have some influence over how that data are used. If a department collects gender data and is interested in examining gender gaps, PER faculty should advocate for best practices in interpreting that data.

Finally, physicists do not teach with the aspiration that their students should be able to score higher on multiple-choice tests. Conceptual inventories are valuable as one measure of learning, but if faculty are checking their data for gender gaps, it is equally worth interrogating other aspects of the classroom. For example, instructors might also benefit from taking an Implicit Attitudes Test to check for their own gender biases. Peer teaching observations can include noting whether the instructor differentially calls on students by gender, and pass or drop rates can likewise be checked for gender gaps. Concept inventories, by their structure, position learning as an individual outcome that is entirely located in the students. This is a useful approximation insofar as it reflects what a single person will carry forward from the classroom. However, that focus on the individual and not on the learning environment also filters out other possible sources of a gender gap. Faculty who find the question valuable enough to ask in one context ("How do my students differ in their scores by gender?") should consider how it fits into a coherent plan to evaluate the gender dynamics in their classroom, and what remediation strategies exist for other aspects of classroom culture.

This research demonstrated the need to use the fair conceptual instruments proposed by Traxler *et al.* [33] and Henderson *et al.* [40], particularly the fair FCI. This research also showed that a substantial part of the gender differences in each sample could be explained by prior preparation. If δG_{equal} is eventually shown to also result from prior preparation, the majority of gender differences in all samples were the result of differences in prior preparation in physics. This suggests that physics classes must deploy instructional strategies to address these differences. These may include adaptive conceptual training that allows all students to work toward a mastery goal, rather than delivering the same assignments to all students, thus giving more conceptual practice to all students who need it.

The partition of the gender gap presented above also has serious implications for future research and the interpretation of past research. Prior preparation differences explained

a substantial part of the gender differences in most samples; academic performance differences explained smaller, but still significant amounts of the differences in some samples. The amount these two factors contributed to the gender gap varied greatly between samples. As such, researchers investigating differences in student performance for any reason must collect an appropriate set of control variables including standardized test scores and measures of student prior preparation. The results of this work also seem to indicate that conceptual pretest scores may not provide an accurate characterization of physics prior preparation; more accurate measures should be developed.

VIII. FUTURE

The gender gap has been an active area of research for over a decade; it seems unacceptable to have so much of the gender differences still unexplained. We intend to refine our measurement of prior preparation with the inclusion of a broad set of high school course-taking measurements to determine how much of δG_{equal} can be explained by high school physics and mathematics preparation.

IX. CONCLUSION

This work partitioned the gender difference in post-test performance on the FCI, FMCE, and CSEM into four components: academic performance, instrumental fairness, physics-specific preparation, and a fourth segment representing other effects. The percentage of the gender gap accounted for by each segment varied strongly between the five samples. Fairness accounted for 30% of the gender differences in the FCI, 17% and 9% of differences in the FMCE, and 2%–6% of differences in the CSEM. For four of the five samples, differences in prior preparation measured by pretest scores accounted for approximately 40% of the gender gap. The amount of the gender gap which was accounted for by other effects varied widely between samples. Further correcting for prior preparation using the post-test score in the previous class reduced the size of the gender differences resulting from other effects, sometimes dramatically. This suggests that a CSEM pretest score does not completely capture the effect of prior preparation on conceptual performance.

ACKNOWLEDGMENTS

This work was supported in part by the National Science Foundation as part of the evaluation of improved learning for the Physics Teacher Education Coalition, PHY-0108787, and by Grants No. EPS-1003907 and No. ECR-1561517. We would also like to thank Steven Pollock for his contribution of one of the samples.

[1] D. Hestenes, M. Wells, and G. Swackhamer, Force Concept Inventory, Phys. Teach. **30**, 141 (1992).
[2] 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).
[3] D. P. Maloney, T. L. O'Kuma, C. Hieggelke, and A. Van Huevelen, Surveying students' conceptual knowledge of electricity and magnetism, Am. J. Phys. **69**, S12 (2001).
[4] A. Madsen, S. B. McKagan, and E. Sayre, Gender gap on concept inventories in physics: What is consistent, what is inconsistent, and what factors influence the gap?, Phys. Rev. ST Phys. Educ. Res. **9**, 020121 (2013).
[5] D. Voyer and S. D. Voyer, Gender differences in scholastic achievement: A meta-analysis, Psychol. Bull. **140**, 1174 (2014).
[6] N. S. Cole, *The ETS Gender Study: How Females and Males Perform in Educational Settings* (Educational Testing Service, Princeton, NJ, 1997).
[7] Y. Maeda and S. Y. Yoon, A meta-analysis on gender differences in mental rotation ability measured by the Purdue spatial visualization tests: Visualization of rotations (PSVT: R), Educ. Psychol. Rev. **25**, 69 (2013).
[8] D. F. Halpern, *Sex Differences in Cognitive Abilities*, 4th ed. (Psychology Press, Francis & Tayler Group, New York, NY, 2012).
[9] J. S. Hyde and M. C. Linn, Gender differences in verbal ability: A meta-analysis, Psychol. Bull. **104**, 53 (1988).
[10] J. S. Hyde, E. Fennema, and S. J. Lamon, Gender differences in mathematics performance: A meta-analysis, Psychol. Bull. **107**, 139 (1990).
[11] J. V. Mallow and S. L. Greenburg, Science anxiety: Causes and remedies, J. Coll. Sci. Teach. **11**, 356 (1982).
[12] M. K. Udo, G. P. Ramsey, and J. V. Mallow, Science anxiety and gender in students taking general education science courses, J. Sci. Educ. Technol. **13**, 435 (2004).
[13] J. Mallow, H. Kastrup, F. B. Bryant, N. Hislop, R. Shefner, and M. Udo, Science anxiety, science attitudes, and gender: Interviews from a binational study, J. Sci. Educ. Technol. **19**, 356 (2010).
[14] N. M. Else-Quest, J. S. Hyde, and M. C. Linn, Cross-national patterns of gender differences in mathematics: A meta-analysis, Psychol. Bull. **136**, 103 (2010).
[15] X. Ma, A meta-analysis of the relationship between anxiety toward mathematics and achievement in mathematics, J. Res. Math. Educ. **30**, 520 (1999).

[16] J. R. Shapiro and A. M. Williams, The role of stereotype threats in undermining girls' and women's performance and interest in STEM fields, Sex Roles **66**, 175 (2012).
[17] L. E. Kost, S. J. Pollock, and N. D. Finkelstein, Unpacking gender differences in students' perceived experiences in introductory physics, AIP Conf. Proc. **1179**, 177 (2009).
[18] A. Miyake, L. E. Kost-Smith, N. D. Finkelstein, S. J. Pollock, G. L. Cohen, and T. A. Ito, Reducing the gender achievement gap in college science: A classroom study of values affirmation, Science **330**, 1234 (2010).
[19] M. Lorenzo, C. H. Crouch, and E. Mazur, Reducing the gender gap in the physics classroom, Am. J. Phys. **74**, 118 (2006).
[20] L. E. Kost, S. J. Pollock, and N. D. Finkelstein, Characterizing the gender gap in introductory physics, Phys. Rev. ST Phys. Educ. Res. **5**, 010101 (2009).
[21] P. B. Kohl and H. V. Kuo, Introductory physics gender gaps: Pre-and post-studio transition, AIP Conf. Proc. **1179**, 173 (2009).
[22] 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).
[23] M. J. Cahill, K. M. Hynes, R. Trousil, L. A. Brooks, M. A. McDaniel, M. Repice, J. Zhao, and R. F. Frey, Multiyear, multi-instructor evaluation of a large-class interactive-engagement curriculum, Phys. Rev. ST Phys. Educ. Res. **10**, 020101 (2014).
[24] N. I. Karim, A. Maries, and C. Singh, Do evidence-based active-engagement courses reduce the gender gap in introductory physics?, Eur. J. Phys. **39**, 025701 (2018).
[25] R. Henderson, G. Stewart, J. Stewart, L. Michaluk, and A. Traxler, Exploring the gender gap in the Conceptual Survey of Electricity and Magnetism, Phys. Rev. Phys. Educ. Res. **13**, 020114 (2017).
[26] G. A. Morris, L. Branum-Martin, N. Harshman, S. D. Baker, E. Mazur, S. Dutta, T. Mzoughi, and V. McCauley, Testing the test: Item response curves and test quality, Am. J. Phys. **74**, 449 (2006).
[27] J. Wang and L. Bao, Analyzing Force Concept Inventory with item response theory, Am. J. Phys. **78**, 1064 (2010).
[28] G. A. Morris, N. Harshman, L. Branum-Martin, E. Mazur, T. Mzoughi, and S. D. Baker, An item response curves analysis of the Force Concept Inventory, Am. J. Phys. **80**, 825 (2012).
[29] M. Planinic, L. Ivanjek, and A. Susac, Rasch model based analysis of the Force Concept Inventory, Phys. Rev. ST Phys. Educ. Res. **6**, 010103 (2010).
[30] S. Osborn Popp, D. Meltzer, and M. C. Megowan-Romanowicz, Is the Force Concept Inventory biased? Investigating differential item functioning on a test of conceptual learning in physics, in *2011 American Educational Research Association Conference* (American Education Research Association, Washington, DC, 2011).
[31] T. F. Scott and D. Schumayer, Students' proficiency scores within multitrait item response theory, Phys. Rev. ST Phys. Educ. Res. **11**, 020134 (2015).
[32] R. D. Dietz, R. H. Pearson, M. R. Semak, and C. W. Willis, Gender bias in the Force Concept Inventory?, AIP Conf. Proc. **1413**, 171 (2012).
[33] A. Traxler, R. Henderson, J. Stewart, G. Stewart, A. Papak, and R. Lindell, Gender fairness within the Force Concept Inventory, Phys. Rev. Phys. Educ. Res. **14**, 010103 (2018).
[34] T. I. Smith and M. C. Wittmann, Applying a resources framework to analysis of the Force and Motion Conceptual Evaluation, Phys. Rev. ST Phys. Educ. Res. **4**, 020101 (2008).
[35] R. M. Talbot, Taking an item-level approach to measuring change with the Force and Motion Conceptual Evaluation: An application of item response theory, School Sci. Math. **113**, 356 (2013).
[36] M. Ishimoto, R. K. Thornton, and D. R. Sokoloff, Validating the Japanese translation of the Force and Motion Conceptual Evaluation and comparing performance levels of American and Japanese students, Phys. Rev. ST Phys. Educ. Res. **10**, 020114 (2014).
[37] T. I. Smith, M. C. Wittmann, and T. Carter, Applying model analysis to a resource-based analysis of the Force and Motion Conceptual Evaluation, Phys. Rev. ST Phys. Educ. Res. **10**, 020102 (2014).
[38] D. E. Meltzer, Analysis of shifts in students' reasoning regarding electric field and potential concepts, AIP Conf. Proc. **883**, 177 (2007).
[39] J. Leppävirta, The effect of naïve ideas on students' reasoning about electricity and magnetism, Res. Sci. Educ. **42**, 753 (2012).
[40] R. Henderson, P. Miller, J. Stewart, A. Traxler, and R. Lindell, Item-level gender fairness in the Force and Motion Conceptual Evaluation and the Conceptual Survey of Electricity and Magnetism, Phys. Rev. Phys. Educ. Res. **14**, 020103 (2018).
[41] A. L. Traxler, X. C. Cid, J. Blue, and R. Barthelemy, Enriching gender in physics education research: A binary past and a complex future, Phys. Rev. Phys. Educ. Res. **12**, 020114 (2016).
[42] R. K. Thornton, D. Kuhl, K. Cummings, and J. Marx, Comparing the force and motion conceptual evaluation and the force concept inventory, Phys. Rev. ST Phys. Educ. Res. **5**, 010105 (2009).
[43] US News & World Report: Education, https://premium.usnews.com/best-colleges.
[44] L. C. McDermott, P. S. Shaffer, and University of Washington Physics Education Group, *Tutorials in Introductory Physics* (Prentice Hall, Englewood Cliffs, NJ, 1998).
[45] V. Otero, S. Pollock, and N. Finkelstein, A physics department's role in preparing physics teachers: The Colorado Learning Assistant Model, Am. J. Phys. **78**, 1218 (2010).
[46] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevPhysEducRes.15.010131 for regression analysis for all samples, and comparison of test average and ACT/SAT percentage as a measure of general academic performance.
[47] R Core Team, *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria 2017).
[48] M. Y. Jaber, *Learning Curves: Theory, Models, and Applications* (CRC Press, Tayler & Francis Group, New York, NY, 2016).

[49] A. Baddeley, *Essentials of Human Memory* (Psychology Press, Taylor & Francis Group, New York, NY, 2014).

[50] S. B. Hofer, T. D. Mrsic-Flogel, T. Bonhoeffer, and M. Hübener, Experience leaves a lasting structural trace in cortical circuits, Nature (London) **457**, 313 (2009).

[51] L. Ding, Theoretical perspectives of quantitative physics education research, Phys. Rev. Phys. Educ. Res. (to be published).

[52] P. Eaton and S. D. Willoughby, Confirmatory factor analysis applied to the force concept inventory, Phys. Rev. Phys. Educ. Res. **14**, 010124 (2018).

[53] L. Ding and X. Liu, Getting started with quantitative methods in physics education research, in *Getting Started in Physics Education Research*, Reviews in PER, Vol. 2, edited by C. Henderson and K. A. Harper (American Association of Physics Teachers, College Park, MD, 2012).

[54] A. Madsen, S. B. McKagan, and E. C. Sayre, Resource letter RBAI-1: Research-based assessment instruments in physics and astronomy, Am. J. Phys. **85**, 245 (2017).