Item-level gender fairness in the Force and Motion Conceptual Evaluation and the Conceptual Survey of Electricity and Magnetism

Rachel Henderson,[1,*] Paul Miller,[1] John Stewart,[1] Adrienne Traxler,[2] and Rebecca Lindell[3]

[1]*West Virginia University, Department of Physics and Astronomy, Morgantown, West Virginia 26506, USA*

[2]*Wright State University, Department of Physics, Dayton, Ohio 45435, USA*

[3]*Tiliadal STEM Education Solutions, Lafayette, Indiana 47901, USA*

(Received 19 February 2018; published 11 July 2018)

Gender gaps on the most widely used conceptual inventories created by physics education researchers have been extensively studied. Most of the research exploring the consistent gender gaps has been performed at the student level using the total evaluation score; less research has been performed examining these assessments at the item level and this research has been predominately restricted to the Force Concept Inventory (FCI). Many studies have identified subsets of FCI items as unfair to either men or women. An item is fair if men and women of equal ability in conceptual physics score equally on the item. This study explored the item-level gender fairness of the Force and Motion Conceptual Evaluation (FMCE) and the Conceptual Survey of Electricity and Magnetism (CSEM). Classical test theory and differential item functioning (DIF) analysis were employed to examine item fairness. Fairness was investigated with four large post-test samples, two for the FMCE ($n_1 = 3016$, $n_2 = 3719$) and two for the CSEM ($n_1 = 2014$, $n_2 = 2657$). Men and women performed significantly differently on the majority of FMCE items but with no more than a small effect size. There were fewer items in the CSEM where men and women performed differently. Using DIF analysis, which assumes that overall test score is an accurate measure of ability, only one item in the FMCE demonstrated large DIF in both samples with that item unfair to women. One additional item showed large DIF in a single sample, also unfair to women. Only one item in the CSEM demonstrated large DIF. The item was unfair to men but this result was not consistent across all samples. The number of large DIF items identified in both the FMCE and the CSEM was substantially smaller than the number of large DIF items identified in the FCI by previous studies.

DOI: 10.1103/PhysRevPhysEducRes.14.020103

I. INTRODUCTION

The properties and performance of the most commonly used conceptual inventories constructed by physics education research (PER) have been studied through factor analysis [1–3], item response theory [4–7], and network analysis [8]. Most of these studies, however, have been performed using the Force Concept Inventory (FCI) [9]; substantially less research has been performed exploring the structure and validity of the Force and Motion Conceptual Evaluation (FMCE) [10] or the Conceptual Survey of Electricity and Magnetism (CSEM) [11].

This work examines the validity and fairness of the FMCE and CSEM using four large samples drawn from calculus-based college physics classes. We adopt the validation framework proposed by Jorion *et al.* [12] for evaluating engineering conceptual inventories. This framework begins with an examination of classical test theory (CTT) difficulty and discrimination to identify items outside of the suggested range on these measures; these items pose reliability and validity problems for the instrument. Item response theory (IRT) is then applied to further understand item functioning. Reliability is assessed with Cronbach's α and inter-item correlations; factor analysis is then applied to understand subscale reliability. The primary focus of this work is to understand gender differences. While the reliability analysis may provide information about overall instrumental validity, Traxler *et al.* [13] found it provided little explanation of the gender differences; we will leave a general reliability analysis of the FMCE and CSEM for future studies.

Our prior work extended the framework to include an item fairness analysis [13]. The Educational Testing Service [14,15], the American Educational Research Association, the American Psychological Association, and National Council on Measurement in Education [16] suggest that fairness analysis is a crucial step in instrument

*rjhenderson@mix.wvu.edu

construction and, further, suggest differential item functioning (DIF) analysis as one part of fairness analysis. Item and instrument fairness is a sometimes contentious topic [17]. We will adopt a narrow definition of fairness: an item will be considered fair if it demonstrates negligible DIF; that is, if the item performs identically for two groups of students with equal ability with the material tested. This work uses many terms that are common within the test development literature that have different meanings than their common usage; we will define these terms as they are introduced. For a careful exploration of the terms and for standards of test development practice see Ref. [16].

Traxler *et al.* examined the item-level fairness and test construction fairness of the FCI [9] with a large data set ($N = 5769$) drawn from four institutions [13]. In the different samples examined, many items fell outside the suggested range for CTT difficulty and discrimination. These "problematic" items were different for men and women. This work also showed that, when the average score for each item was plotted for men and women, five items stood out as substantially unfair to women while most other items were somewhat unfair to women. DIF analysis confirmed the unfairness of these five items and identified a total of 8 unfair items with large DIF; 6 unfair to women and 2 unfair to men. Most of these items had also been identified as unfair in previous studies, but not consistently [18,19]. An unbiased FCI was then constructed by iteratively removing unfair items which produced a 20-item instrument including only fair items. Beyond the 8 items initially identified with large DIF, one additional large DIF and one small-to-moderate DIF item were uncovered as the large DIF items were removed from the instrument. This instrument was further reduced by removing item 29 which consistently failed reliability and validity metrics. The gender gap on the 19-item FCI was reduced by nearly 50% in the primary sample. The current work partially replicates this analysis path for the FMCE and the CSEM.

While there is a substantial body of research investigating the gender fairness of the FCI, little research into the gender fairness of the FMCE or CSEM exists. Very little work on test construction fairness has been performed on instruments in PER. Traxler *et al.* defined an instrument's test construction as being fair if the instrument had similar reliability and validity properties for all populations. The only similar work we know of was performed by Henderson *et al.* [20] which identified differences in the validity of the CSEM pretest score for men and women. A few developers of concept inventories have noted gender differences in average student score [21,22]. This information was reported at the whole-instrument level and not by item. Engelhardt [23] recommends that developers look for gender effects in their item-level analyses, but it is not yet clear whether this suggestion has been taken up by the PER community.

As with most other studies of gender fairness, this work will treat gender as a binary variable. This treatment obscures the complicated nature of gender identity and future studies should perform a more thorough investigation. For a more detailed discussion of gender in physics see Traxler *et al.* [24].

A. Reliability and validity

Much of the work exploring the reliability, validity, and structure of physics conceptual inventories has been performed on the FCI. These studies have included analyses of the factor structure [1,2,7,25,26], reliability [27,28], and item properties measured with IRT [5] and with item response curves [4].

Thornton *et al.* provided evidence for the validity of the FMCE by comparing FMCE post-test scores with FCI post-test scores and found a strong correlation [29]. Only one study examined the factor structure of the FMCE. Ramlo found the factor structure of the FMCE pretest was undefined but that a three-factor solution existed for the FMCE post-test [30]. To our knowledge, little research has been performed investigating the reliability, validity, or fairness of either the FMCE or the CSEM.

B. The "gender gap"

The overall gender gap on the various introductory-level conceptual inventories used in PER has been extensively studied. Madsen, McKagan, and Sayre summarized reported performance differences on conceptual evaluations between men and women [31]. Men, on average, outperform women on the mechanics conceptual inventories by 13% on the pretest and by 12% on the post-test. On electricity and magnetism conceptual inventories, men also typically outperform women, with men scoring 3.7% higher on the pretest scores, 8.5% on the post-test.

One promising subset of gender gap studies focuses on the role that instructors and physics education researchers can play, by examining the effects of reformed pedagogy. Although some studies have shown that interactive engagement techniques decrease the overall gender gap [32,33], these results have not been consistent [34–36].

The following subsections review the exploration of gender gaps and their causes in the FCI, the FMCE, and the CSEM. For an overview of the research performed on the gender gap in physics, see Madsen, McKagan, and Sayre [31]. For a more detailed discussion of the possible sources behind the consistent gender gaps in physics conceptual inventories, and for related work in higher education, see Henderson *et al.* [20].

1. Gender and the FCI

Most work exploring gender differences in physics conceptual performance has been conducted with the FCI. Studies have explored the relation of gender and post-test scores with other correlates such as scientific reasoning [37,38], standardized test scores [39], other

pretest scores [40,41], and psychological factors such as self-efficacy [42,43]. See Traxler *et al.* for a more complete review [13].

A few studies have sought alternate perspectives on gendered responses to FCI questions. McCullough [44] found that switching the gender context from stereotypically masculine scenarios (hockey, rockets, etc.) to stereotypically feminine contexts significantly changed the gender gap on a number of items. McCaskey and collaborators [42,43] asked students to mark both their own beliefs and what a scientist would say, and found that women showed more "splits" between the two sets of answers. No similar work that we know of has been done with the FMCE or CSEM, and these three studies all used data outside the most common PER setting of calculus-based courses. The analysis in this paper will pursue the more traditional psychometric methods of CTT and DIF, but these alternate ways of probing the structure of concept inventories may hold important clues for future research.

2. Gender and the FMCE

The FMCE has been used to measure student conceptual understanding of Newton's laws of motion for nearly 20 years. Pollock, Finkelstein, and Kost investigated the effect of interactive-engagement techniques on the difference in performance on the FMCE between men and women [33]. Even though previous research showed that the gender gap on the mechanics inventories could be reduced by using these techniques [45], the results of the Pollock, Finkelstein and Kost study did not support this finding; interactive engagement was not sufficient to reduce the differences in performance on the FMCE.

Kost, Pollock, and Finkelstein explored factors that contributed to the gender gap in the FMCE post-test [45] including background and preparation differences, students' attitudes toward science measured by the Colorado Learning Attitudes about Science Survey (CLASS) [46], and other assessments such as FMCE pretest scores and math placement exam scores. Student background and preparation differences explained a substantial amount of the variance in the gender gap on the FMCE post-test. We will "bin" FMCE post-test score by pretest, similar to the Kost *et al.* [45] study. A bin is defined as a range of pretest scores. In a different study, the same authors investigated the effect of physics identity and self-efficacy on student performance [47]. Neither physics identity nor self-efficacy explained the gender gap in FMCE post-test scores.

3. Gender and the CSEM

The gender gap on the electricity and magnetism conceptual inventories, such as the CSEM, has received significantly less attention than both the FCI and the FMCE. Although Madsen, McKagan, and Sayre report that, on average, men outperform women on both the CSEM pretest and post-test, the gender gap on the CSEM has been less consistent. While most studies report a male advantage, one study reported women outperforming men postinstruction [48].

Kohl and Kuo compared differences in normalized gain between men and women as a function of binned CSEM pretest score [36]. In three of the four bins, a significant gender gap in normalized gains was measured; however, the gender gap was not significant for the bin containing raw scores of 0 to 4 out of 32.

Henderson *et al.*, using a similar binning by pretest score, found an overall gender gap in the CSEM pretest and post-test. This gap was also present in other qualitative assessment items such as qualitative lab quiz problems and qualitative in-semester examination problems, but not in quantitative exam problems [20]. The gender differences in each of the qualitative problem sets grew as the students' CSEM pretest score increased. As in the Kohl and Kuo study, no gender gap was measured in the lowest CSEM pretest bins (raw scores between 0 and 8). Henderson *et al.* suggested that, because no gender gap was measured for students scoring below a 25% on the pretest, the CSEM was not intrinsically biased. Henderson *et al.* argued that the CSEM pretest provided a less accurate measure of the incoming physics conceptual knowledge of women compared to men; the CSEM was less valid for women than for men when used as a pretest.

C. Item analysis

Within the Jorion framework, instrument validation begins with item analysis which determines if CTT difficulty P and discrimination D are within an established range. For distractor-driven instruments, Jorion suggests well-functioning items have $D > 0.2$ and $0.2 < P < 0.8$ [12].

1. FMCE

Although most of the research on the FMCE examined overall scores pre- and postinstruction, some studies have investigated individual items on the FMCE. Talbot investigated the change in Newtonian thinking at the item level, arguing that this would give more detailed insight into student understanding of Newtonian mechanics [49]. Items 36 and 38 were too difficult ($P < 0.2$) on the pretest and items 40, 41, 42, and 43 were too easy ($P > 0.8$) on both the pretest and the post-test.

In a study comparing the performance of Japanese students to American students on the FMCE, each of the FMCE items was translated to Japanese [50]. CTT item difficulty P and item discrimination D were analyzed for the Japanese students showing that the majority of the FMCE items fell in the range of the desired difficulty. In addition, items 36 and 38 were classified as difficult items and items 40 through 43 were identified as easy items, which was consistent with the study performed by Talbot. Because the difficulty and discrimination were similar to

those of the American students, the authors concluded that the FMCE could be used to compare American and Japanese students.

While performing a comparison between FCI and the FMCE, Thornton *et al.* classified certain groups of items on the FMCE as "distinct clusters" [29]. For example, the three items assessing student understanding of acceleration of a tossed coin (items 27, 28, and 29) were defined as one cluster, 27_29. The notation 27_29 indicates the group of items 27, 28, and 29. Clustered problems are a set of problems that measure the same concept. Thornton [29] suggested that a cluster be graded as correct only if a student answered all questions in the cluster correctly.

The three distinct clusters described by Thornton *et al.* have been studied at the item level [51,52]. In 2008, Smith and Wittmann introduced revised clusters and investigated student response patterns on those sets of items. This work suggested that the two distinct clusters defined by Thornton *et al.*, 8_13 and 27_29, should be combined into one cluster described as *reversing direction*. They also introduced cluster 40_43 as *velocity graphs* [51]. In 2014, Smith, Wittmann, and Carter used these revised clusters to provide insight into the effectiveness of interactive classroom techniques [52].

Overall, the analyses that have investigated the FMCE at the item level treat the students as an undifferentiated sample; the comparison of CTT difficulty and discrimination between men and women for the FMCE has not yet been reported.

2. CSEM

There have been very few studies that have focused on the individual items of the CSEM. Maloney *et al.* reported that the difficulty of the items on the CSEM were between 0.1 and 0.9 [11]. This analysis was performed for both the algebra-based and calculus-based introductory electricity and magnetism courses. Only one item, item 3, had a difficulty of above 0.8 and three items seemed to be too challenging with a difficulty of less than 0.2 (items 14, 20, and 31). Item discrimination was also evaluated; only four items had a discrimination less than 0.2; however, the authors do not specify which items.

Some studies have conducted analyses on a few specific items on the CSEM. Meltzer explored the shifts from pretest to post-test in student responses and reasoning on items 18 and 20, which ask students to compare the magnitude and direction of electric field and electric force, respectively, at two different points on equipotential lines [53]. Leppävirta investigated students' alternate ideas on the items that assess student understanding of Newton's 3rd law on the CSEM (items 4, 5, 7, and 24) [54]. One out of five students had an alternate model of Newton's 3rd law prior to electricity and magnetism instruction; however, postinstruction these students are likely to change their understanding to the correct model.

To our knowledge, no item level analysis of the CSEM has been reported differentiated by gender.

D. Item fairness analysis

Multiple studies have investigated item fairness within the FCI. Some studies have examined performance differences between genders on the FCI at the item level. Dietz *et al.* investigated a balanced sample of men and women and found that items 4 and 9 were unfair to men and item 23 was unfair to women [18]. Osborn Popp, Meltzer, and Megowan-Romanowicz analyzed the FCI for high school physics students and showed fourteen items that were significantly unfair, but only item 23 showed large DIF while items 4, 6, 9, 14, 15, and 29 had small to moderate DIF [19]. More recently, Traxler *et al.* investigated item-level gender fairness in the FCI and found eight items that were substantially unfair, six unfair to women and two unfair to men [13].

A review of the literature did not identify any studies investigating the fairness of either the FMCE or CSEM.

E. Research questions

This study seeks to answer the following research questions:

RQ1: Are there items in the FMCE or the CSEM which CTT would identify as problematic? Are the problematic items the same for men and women?

RQ2: Are there items in the FMCE or the CSEM which are substantially unfair to men or women?

RQ3: Are the differences in overall performance between men and women on the FMCE or the CSEM dependent on the student's FMCE or CSEM pretest score?

II. METHODS

A. Instruments

The FMCE is a 43-item conceptual inventory evaluating students' conceptual understanding of Newton's laws of motion [10]. The assessment uses extensive blocking of items referencing common physical systems to probe students' views of force and motion concepts. Such systems include, but are not limited to, "force sled" questions, "cart on ramp" questions, "coin toss" questions, and "force graph" questions. For each item in a block, there are at least 6 possible responses, some of which were constructed to match students' common misconceptions about force and motion. A revised version was published and includes four additional questions on energy concepts; however, typically these items are not included in the scoring of the FMCE. The second version of the FMCE is available at PhysPort [55].

The CSEM is a 32-item inventory evaluating students' conceptual understanding of electricity and magnetism [11]. Maloney *et al.* developed the CSEM based upon the list of concepts that were initially constructed by Hieggelke and O'Kuma from two preliminary versions measuring conceptual understanding of electricity and magnetism separately [56]. After many iterations along with open-ended versions to identify common misconceptions, the two separate inventories were combined into one assessment designed to measure electricity and magnetism together. The instrument contains questions on Coulomb's force law, the vector addition of electric force, electric field, and magnetic field, as well as induction. For a list of all of the concepts evaluated by the CSEM, see Maloney *et al.* [11]. The final version of the CSEM is also available at PhysPort [55].

B. Samples

Sample 1: Sample 1 was collected for four semesters in the calculus-based introductory mechanics class at a large western land-grant university in the US serving 34 000 students. The university had a Carnegie classification of "highest research activity" for the period studied [57]. The general undergraduate population had a range of ACT scores from 25–30 (25th to 75th percentile range) [58]. The general undergraduate population had a demographic composition of 69% White, 11% Hispanic, 7% International, 5% Asian, and 5% two or more races with all other groups with representation of less than 5% [58].

The course was taught by four faculty members and shared a common format throughout each semester. Each week, the course consisted of three 50 min lectures and one 50 min tutorial section where the University of Washington *Tutorials in Introductory Physics* [59] were led by a graduate teaching assistant and an undergraduate learning assistant. Lecture instructors used peer instruction with clickers. Students were assigned weekly homework as well as prelecture videos. Students were assessed with three in-semester examinations and a final examination. The FMCE pretest and post-test were administered during the tutorial section; while attendance was required, the pretest and post-test did not count toward the student's final grade. No laboratory was associated with the course. The aggregated sample consisted of 3511 FMCE pretest responses (74% men) and 3016 FMCE post-test responses (73% men); there were 2744 matched pretest and post-test pairs.

Sample 2: Sample 2 was collected for a total of 14 semesters in the calculus-based electricity and magnetism course at a large southern land-grant university serving approximately 25 000 students. The general undergraduate population had a range of ACT scores from 23 to 29 (25th to 75th percentile range) [58]. The university had a classification of highest research activity for the entire period studied [57]. The overall undergraduate demographics were 77% White, 8% Hispanic, 5% African American, 2% Asian with other groups each 3% or less [58].

The course was taught and overseen by one lead instructor over the time period studied. The course consisted of two 50 min lectures and two 2 h laboratory sessions each week. Students completed four in-semester examinations, weekly homework assignments, in-class lecture quizzes, and laboratory quizzes. The CSEM was given as a laboratory quiz pre- and postinstruction. The score on the CSEM was counted toward the students' course grade. The aggregated data set ($n_{\mathrm{pre}} = 2108$, $n_{\mathrm{post}} = 2014$) consisted of only students who completed the course for a grade and received credit for both the CSEM pretest and the CSEM post-test. The sample was primarily male (77%). There were 1804 matched pretest and post-test pairs.

Sample 3: Sample 3 was collected during 13 semesters from Spring 2011 to Spring 2017 at a large eastern land-grant university serving approximately 30 000 students. In 2016, this institution first achieved a Carnegie classification of highest research activity [57]. The undergraduate ACT range for this institution was 21–26 (25th to 75th percentile range). The overall undergraduate demographics were 79% White, 6% International, 5% African American, 4% Hispanic, 2% Asian with other groups each 4% or less [58]. The students in sample 3 were enrolled in either the introductory, calculus-based mechanics course (sample 3A) or the introductory, calculus-based electricity and magnetism course (sample 3B). Only the students who completed the courses for a grade and completed both the pretest and post-test were included. The Sample 3A data set included 3956 pretest responses (80% men) and 3719 post-test responses (80% men) from the FMCE. There were 3719 matched pretest–post-test pairs in sample 3A. The sample 3B data set included 3185 pretest responses (83% men) and 2,657 post-test responses (81% men) from the CSEM. There were 2439 matched pretest and post-test pairs in sample 3B.

The instructional environment for sample 3 was quite variable for the period studied and may, therefore, be representative of a sample drawn from multiple institutions with the same student characteristics. Between Spring 2011 and Spring 2015 semesters, a Learning Assistant (LA) program [60] was implemented as a tool to improve conceptual understanding of students in the introductory calculus-based sequence. During this time, the students attended four 50 min lectures and one 2 h laboratory each week. Over this period, the lectures were presented by 14 different instructors of various standing from full professors to late career graduate students. Many of these instructors taught multiple lecture sections and were able to establish their own teaching pedagogy, homework assignments, and examination policies. In the laboratory, the first hour was dedicated to students working in small groups on the University of Washington *Tutorials in Introductory Physics*

[59] with the LA acting as the lead lab instructor. The LA received training from both an expert in science education and from an experienced physics instructor. In the second hour of lab, the students worked on a traditional laboratory experiment with the graduate teaching assistant (TA) acting as the lead lab instructor. Lab reports and short homework assignments from the *Tutorials in Introductory Physics* were collected and graded by the TA.

The LA program was discontinued after the Spring 2015 semester because it had reached the end of its funding. After the LA program, between the Fall 2015 and Spring 2017 semesters, each course was team taught by a pair of experienced educators. The courses consisted of three 50 min lectures and one 3 h laboratory. All sections of this course used the same in-class examination policies and similar homework policies. All lecture sections employed clickers to engage students in conceptual learning. Credit for the completion of the FMCE was given for a good faith effort and credit for the completion of the CSEM was dependent on the instructor.

Samples 3A and 3B aggregate data over a number of instructional environments. The analysis was repeated separately for the period Spring 2011 to Spring 2015 and the period Fall 2015 to Spring 2017; the conclusions were the same for each period.

C. FMCE scoring

A modified scoring method for the FMCE proposed by Thornton *et al.* was employed in this study [29]. A composite score of the original FMCE 43 items is formed to produce a score out of 33 possible points. Items 5, 15, 33, 35, 37, and 39 were eliminated because students could "expertly" answer these items prior to becoming a consistent Newtonian thinker [10,61]. Item 6 was also eliminated because physics experts frequently answered this item incorrectly.

In addition to eliminating these items, Thornton *et al.* proposed an "all-or-nothing" scoring method for the three clusters of items examining acceleration (items 8_10, 11_13, and 27_29). The authors argued that a student does not fully understand the concept of acceleration unless he or she answers all three parts of the cluster correctly. For students who do answer all three parts correctly, two points are given toward their overall score and zero points otherwise.

In our analysis, the method of all-or-nothing scoring system was employed; however, only one point was rewarded to the students who answered each of the three parts correctly. With the elimination of the 7 items and the modified all-or-nothing scoring method, the students' FMCE score was out of 30 possible points. This modification was made to conform with the requirements of DIF analysis (i.e., the assumption that all items are equally weighted).

D. Classical test theory

CTT is an important component of modern measurement theory [62]. Ding and Beichner summarize five approaches to analyzing multiple-choice questions including CTT [63]. The current study will use CTT item difficulty and discrimination measures. In a previous study, CTT difficulty and discrimination were presented in parallel with their IRT counterparts; the two methods gave generally consistent results [13].

Item difficulty P measures how "easy" an item is for students. It is defined as the proportion of correct responses for a given population, the average score on the item (the higher the item difficulty, the easier the item) [64]. Item discrimination D measures how well an item can distinguish students who have strong knowledge of the subject matter from those who do not. Discrimination is defined as

$$D = P_u - P_l, \tag{1}$$

where P_u is the proportion of participants in the top 27% of the total score distribution answering the question correctly and P_l is the proportion of participants in the bottom 27% answering the item correctly [64].

An item with difficulty or discrimination that are either too high or too low can provide inaccurate information about the population; such items are called "problematic." Jorion suggests items with $D < 0.2$, $P < 0.2$, or $P > 0.8$ as problematic for distractor-driven instruments [12,65,66].

CTT and IRT were also compared for the data sets used in this study. For the FMCE samples, samples 1 and 3A, results were similar with no items standing out as substantially unfair. For the CSEM samples, sample 2 and 3B, the IRT analysis produced substantially larger error bars than the CTT analysis making interpretation of the results ambiguous. As such, only CTT difficulty and discrimination will be reported in detail here.

The phi coefficient ϕ is calculated to explore the differences in the CTT item-level difficulty between men and women [67]. For the phi coefficient, $\phi = 0.1$ is considered a small effect, $\phi = 0.3$ a medium effect, and $\phi = 0.5$ a large effect.

The standard deviation of P and D were calculated by bootstrapping. Bootstrapping is a statistical technique used to estimate variation in models by forming subsamples with replacement of the original data set [68]. For this research, 1000 subsamples were used for each standard deviation estimated.

E. Differential item functioning

In an extension of the Jorion framework, Traxler *et al.* [13] explored item-level fairness in the FCI with graphical analysis and using DIF analysis. DIF assumes the total score on the instrument is an accurate measure of ability. We will measure DIF with the Mantel-Haenszel (MH) statistic [69–71] which has been employed by the

TABLE I. Summary of item statistics and effect sizes reported in this study.

Measure	Description	Usage and range notes
P	Item difficulty	Values from 0 (hardest) to 1 (easiest); consider rejecting items with $P < 0.2$ or $P > 0.8$
D	Item discrimination	Values from -1 (least discriminating) to 1 (most); consider rejecting items with $D < 0.2$
d	Cohen's d	Difference in overall pre- or post-test averages: 0.2 small, 0.5 medium, 0.8 large
ϕ	Phi coefficient	Effect size of the difference between P_F and P_M: 0.1 small, 0.3 medium, 0.5 large
$\Delta\alpha_{\mathrm{MH}}$	Mantel-Haenszel	$\lvert\Delta\alpha_{\mathrm{MH}}\rvert < 1$, negligible; [1, 1.5), small to moderate; > 1.5, large

Educational Testing Service (ETS) for 25 years to examine item fairness in high stakes exams [72]. The MH statistic uses the total score on the instrument to divide the students into groups and then calculates a common odds ratio, α^i_{MH}, comparing the odds of answering an item i correctly for women to the odds of answering an item i correctly for men [73]. The assumption that overall test score is a good measure of ability might be problematic if the overall score is biased; however, the purpose of the MH statistic is to detect differences in item performance, not overall instrument fairness. DIF detects items that stand out as unfair; removing these items is the first step in producing a fair instrument. Once the items that stand out as unfair are removed, general instrumental fairness still should be established by additional analysis.

The α_{MH} statistic can be transformed into an effect size, $\Delta\alpha_{\mathrm{MH}}$, defined by $\Delta\alpha_{\mathrm{MH}} = -2.35\ln(\alpha_{\mathrm{MH}})$ [72]. In this study, men have an advantage when $\Delta\alpha_{\mathrm{MH}} < 0$ and women have an advantage when $\Delta\alpha_{\mathrm{MH}} > 0$. The ETS classifies $|\Delta\alpha_{\mathrm{MH}}| < 1$ as negligible DIF, $1 \leq |\Delta\alpha_{\mathrm{MH}}| < 1.5$ as small to moderate DIF, and $|\Delta\alpha_{\mathrm{MH}}| \geq 1.5$ as large DIF [74]. This classification is called the ETS delta scale.

DIF analysis detects differences in item performance under the assumption that the total instrument score is an accurate measure of each group's proficiency with the material. DIF analysis cannot detect overall instrumental bias; it cannot detect if the majority of items in an instrument favor one group. DIF, then, detects items where the difference in performance on the item is substantially different than the average difference on all items.

F. Other analyses

The differences in performance between men and women were measured with t tests. Cohen's d was used to characterize the effect size for each test; Cohen identified $d = 0.2$ as a small effect, $d = 0.5$ as a medium effect, and $d = 0.8$ as a large effect [75]. Table I provides a summary of the statistics used in this paper.

Because of the number of statistical tests performed in this work, a Bonferroni correction was applied to adjust for the inflation of type I error rate. This correction adjusted the significance levels by dividing the p values by the number of statistical tests performed in the analysis [76].

All statistical analyses were performed with the R software package [77]. DIF analysis was performed with the difR package [78]. Bootstrapping was performed with the boot package [68,79].

III. RESULTS

For each instrument, item fairness will be examined graphically and with DIF analysis. The relation of binned pretest score to post-test score will also be examined.

A. FMCE

Table II presents the FMCE pretest and post-test averages for sample 1 and sample 3A. In sample 1, men outperformed women by 15% on the FMCE pretest and by 14% on the FMCE post-test. These differences were significant for both the FMCE pretest [$t(2059) = 16.69$, $p < 0.001$, $d = 0.57$] and the FMCE post-test [$t(1408) = 12.60$, $p < 0.001$, $d = 0.53$] with medium effect sizes. In sample 3A, significant gender differences were detected on both the FMCE pretest and post-test; however, these differences were smaller than those of sample 1. Men outperformed women by 6% on the pretest [$t(1739) = 16.69$, $p < 0.001$, $d = 0.31$] and by 12% on the post-test [$t(1367) = 11.69$, $p < 0.001$, $d = 0.43$] each with small effect sizes.

1. Item analysis

CTT identifies problematic items as those with difficulty outside of the range from 0.2 to 0.8 ($P < 0.2$ or $P > 0.8$) and those with discrimination less than 0.2 ($D < 0.2$). The problematic items in the FMCE for sample 1 and sample 3A are presented in Table III. For sample 1, nearly half of the items on the FMCE pretest were problematic for women with $P < 0.2$ except for items 40 and 43 with $P > 0.8$. Fewer problematic items were identified for men; items 36 and 38 ($P < 0.2$) and items 40, 42, and 43

TABLE II. FMCE pretest and post-test averages for samples 1 and 3A. Averages are reported as percentages.

		Men		Women	
	N	N	$(M \pm \mathrm{SD})\%$	N	$(M \pm \mathrm{SD})\%$
Sample 1					
FMCE Pretest	3511	2607	45 ± 28	904	30 ± 22
FMCE Post-test	3016	2192	74 ± 26	824	59 ± 28
Sample 3A					
FMCE Pretest	3956	3146	25 ± 19	810	20 ± 14
FMCE Post-test	3719	2947	53 ± 28	772	41 ± 24

TABLE III. CTT problematic items with $P < 0.2$, $P > 0.8$, or $D < 0.2$ for the FMCE.

Gender	Pre or Post	Problematic items
		Sample 1
Women	Pre	2, 4, 8_10, 11_13, 14, 17, 18, 19, 20, 27_29, 36, 38, 40, 43
	Post	40, 43
Men	Pre	36, 38, 40, 42, 43
	Post	22, 24, 26, 31, 40, 41, 42, 43
Overall	Pre	8_10, 36, 38, 40, 42, 43
	Post	15, 24, 26, 40, 41, 42, 43
		Sample 3A
Women	Pre	1, 2, 3, 4, 7, 8_10, 11_13, 14, 16, 17, 18, 19, 20, 21, 23, 25, 27_29, 30, 32, 34, 36, 38, 40, 43
	Post	8_10, 11_13, 40, 43
Men	Pre	1, 2, 4, 8_10, 11_13, 14, 16, 17, 18, 19, 20, 21, 27_29, 30, 32, 34, 36, 38, 40, 43
	Post	40, 42, 43
Overall	Pre	1, 2, 4, 8_10, 11_13, 14, 16, 17, 18, 19, 20, 21, 27_29, 30, 32, 34, 36, 38, 40, 43
	Post	40, 42, 43

($P > 0.8$) were problematic for men. For women in sample 1, only items 40 and 43 were problematic postinstruction ($P > 0.8$). While the number of problematic items decreased for women from pretest to post-test, men had more problematic items after instruction. All of the items identified as problematic for men postinstruction had a difficulty of $P > 0.8$.

The results of the FMCE for sample 3A were fairly similar to those in sample 1. In sample 3A, however, both men and women had many pretest items that were problematic; nearly half of the FMCE pretest items were problematic with $P < 0.2$ for both men and women. As in sample 1, items 40 and 43 were problematic with $P > 0.8$ on the FMCE pretest. The number of problematic items after instruction was reduced for both men and women in sample 3A. Items 40 and 43 continued to be problematic for students on the FMCE post-test. For women, in addition to items 40 and 43, two of the three clustered items identified by Thornton *et al.* [29], 8_10 and 11_13, were problematic ($P < 0.2$). For men, in addition to items 40 and 43, item 42 was a problematic item with $P > 0.8$.

2. Graphical analysis

Item fairness can be explored by plotting the CTT difficulty for men against the CTT difficulty for women as shown in Fig. 1. If men and women have equal proficiency in answering FMCE items, a fair item has the same difficulty for men and women. A line with a slope one, the "fairness line," is also plotted in Fig. 1. A fair item would lie directly on this line. Items that are unfair to women lie above the fairness line while items that are unfair to men lie below the fairness line. Figure 1 shows differences in conceptual performance by gender on the FMCE with the majority of items significantly off the fairness line. The error bars in the figure represent 1 standard deviation in each direction.

The FMCE pretest and post-test results for sample 1 are presented in Figs. 1(a) and 1(b). For sample 1, a chi-squared test showed that for all items in the FMCE pretest, the differences in item difficulties between men and women were significant. The phi coefficient ϕ was calculated for each item to characterize the effect size. Postinstruction, all items except for items 30 and 43 were significantly different for men and women with women scoring lower; however, none of the items showed more than a small effect size.

The FMCE pretest and post-test results for sample 3A are presented in Figs. 1(c) and 1(d). The results were generally similar to the sample 1 results. The sample 3A pretest score was substantially lower than the sample 1 pretest score, which may have produced the clustering near the fairness line at scores less than 25% seen in Fig. 1(c). After instruction, the overall item difficulties for men and women increased; however, most of the items were still significantly different for men and women. Only items 30, 31, 32, 34, 36, and 38 were not significantly different for men and women. None of the items had a difference representing greater than a small effect size.

The figures indicate an overall difference in performance by men and women on the FMCE, an observation that is supported by the significant differences in overall pretest and post-test scores. The plots can also be used to detect differentially functioning items that stand out as substantially more unfair than an average item. Unlike previous work on the FCI [13], there were no set of items that performed significantly differently than most other items. In the FCI, while most items were near the fairness line, five items were visually separate, many standard deviations from the fairness line. The graphical analysis of this section suggests that, at the item level, all of the FCME items function approximately the same for men and women; however, overall, men have a general advantage on the instrument. This analysis cannot determine the origin of the general difference in the performance of men and women on most items, which may result for a number of sources discussed in the introduction from overall instrumental bias to differences in the physics preparation of men and women in the samples.

3. DIF analysis

DIF analysis assumes that students' FMCE post-test score is an accurate measure of their overall ability to

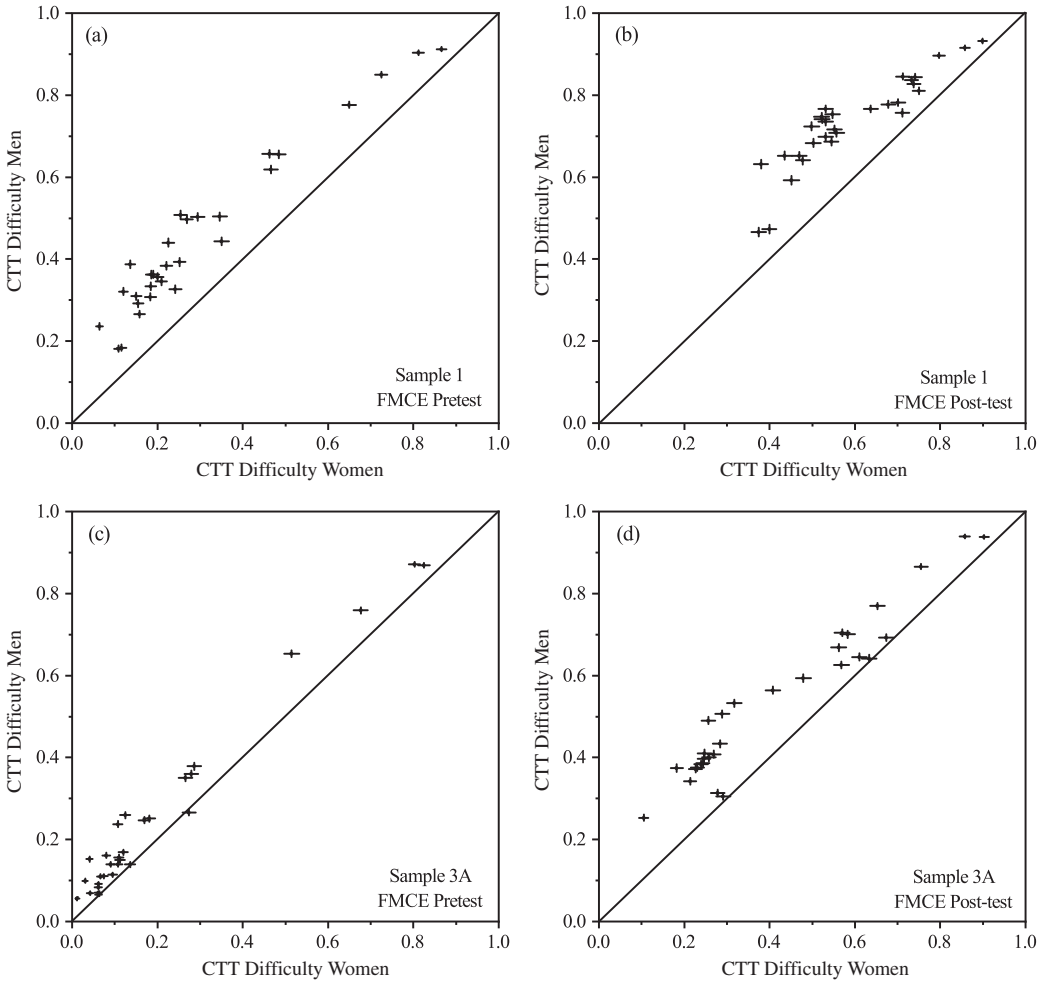

FIG. 1. CTT difficulty results for the FMCE. The top two panels are sample 1 (a) FMCE pretest and (b) FMCE post-test. The bottom two panels are sample 3A (c) FMCE pretest and (d) FMCE post-test. A line of slope 1 is drawn to allow comparison of male and female difficulty. Error bars represent 1 standard deviation in each direction.

answer conceptual physics questions; differences in item performance grouping students by overall test score are characterized by $\Delta\alpha_{\mathrm{MH}}$. In the previous section, which investigated differences in scores graphically, an item was considered fair if the difficulty was equal for men and women. In this section using DIF analysis, the post-test score is assumed to be a accurate measure of the ability of the students. An item will be considered fair if the difference in performance between men and women is equal to the overall difference in post-test score. DIF analysis then identifies items that stand out as unfair against the overall difference in performance of men and women.

Table IV presents the FMCE post-test $\Delta\alpha_{\mathrm{MH}}$ statistic for items in both samples that have either small to moderate or large DIF; these items function differently for men and women taking into account the general difference in post-test score. For sample 1, only one item, item 27_29, demonstrated large DIF with an advantage to men. Item 27_29 is a clustered problem discussing the acceleration of a coin that is tossed straight up into the air. With a value of $\Delta\alpha_{\mathrm{MH}} = -1.50$, this item was on the border between a classification of small to moderate DIF and a classification of large DIF. The other 5 items in sample 1 presented in Table IV (3, 8_10, 11_13, 21, and 30) were measured to have small to moderate DIF with the majority of these items with an advantage to men.

The results for sample 3A were similar; however, in addition to item 27_29, item 40 also had large DIF with an advantage to men. Item 40 is within the force graph block of questions; it involves a car moving toward the right at a constant velocity. Items 3, 7, 8_10, 11_13, 30, 31, 32, 36, 38, and 42 demonstrated small to moderate DIF, half with an advantage to women, half to men.

Because large DIF items influence the overall test score, the identification of DIF can change as problematic items are removed. Large and then small-to-moderate DIF items were iteratively removed from the FMCE and DIF recalculated. For sample 1, items 3, 7, 8_10, 11_13, 21, 25, and 27_29 were removed to produce an instrument with no items with small-to-moderate or large DIF. Eliminating

TABLE IV. CTT difficulty and discrimination and DIF $\Delta\alpha_{MH}$ for the FMCE post-test items with small-to-moderate or large DIF. The significance levels have been Bonferroni corrected: "a" denotes $p < 0.0016$, "b" denotes $p < 0.0003$, and "c" denotes $p < 0.00003$.

Item	P_M	P_F	D_M	D_F	ϕ	$\Delta\alpha_{MH}$
			Sample 1			
3	0.74 ± 0.01	0.52 ± 0.02	0.67 ± 0.02	0.75 ± 0.03	0.21[c]	-1.01[b]
8_10	0.63 ± 0.01	0.38 ± 0.02	0.83 ± 0.02	0.87 ± 0.02	0.23[c]	-1.27[c]
11_13	0.75 ± 0.01	0.52 ± 0.02	0.70 ± 0.02	0.86 ± 0.02	0.22[c]	-1.19[c]
21	0.72 ± 0.01	0.50 ± 0.02	0.70 ± 0.02	0.82 ± 0.03	0.21[c]	-1.09[b]
27_29	0.77 ± 0.01	0.53 ± 0.02	0.65 ± 0.02	0.89 ± 0.02	0.23[c]	-1.50[c]
30	0.76 ± 0.01	0.71 ± 0.02	0.56 ± 0.02	0.50 ± 0.04	0.05	1.05[c]
Item	P_M	P_F	D_M	D_F	ϕ	$\Delta\alpha_{MH}$
			Sample 3A			
3	0.51 ± 0.01	0.29 ± 0.02	0.81 ± 0.01	0.60 ± 0.04	0.18[c]	-1.31[c]
7	0.53 ± 0.01	0.32 ± 0.02	0.76 ± 0.02	0.63 ± 0.04	0.17[c]	-1.25[c]
8_10	0.25 ± 0.01	0.10 ± 0.01	0.74 ± 0.02	0.33 ± 0.04	0.14[c]	-1.16
11_13	0.37 ± 0.01	0.18 ± 0.01	0.83 ± 0.01	0.51 ± 0.04	0.16[c]	-1.25[b]
27_29	0.49 ± 0.01	0.26 ± 0.02	0.84 ± 0.01	0.64 ± 0.03	0.19[c]	-1.66[c]
30	0.64 ± 0.01	0.63 ± 0.02	0.58 ± 0.02	0.50 ± 0.04	0.01	1.09[c]
31	0.69 ± 0.01	0.67 ± 0.02	0.58 ± 0.02	0.53 ± 0.04	0.02	1.03[c]
32	0.65 ± 0.01	0.61 ± 0.02	0.69 ± 0.02	0.68 ± 0.04	0.03	1.10[c]
36	0.30 ± 0.01	0.29 ± 0.02	0.56 ± 0.02	0.51 ± 0.04	0.01	1.39[c]
38	0.31 ± 0.01	0.28 ± 0.02	0.56 ± 0.02	0.52 ± 0.04	0.03	1.03[b]
40	0.94 ± 0.00	0.86 ± 0.01	0.16 ± 0.01	0.32 ± 0.04	0.12[c]	-1.62[c]
42	0.87 ± 0.01	0.76 ± 0.02	0.31 ± 0.02	0.46 ± 0.04	0.12[c]	-1.03[b]

these items reduced the gender gap in FMCE post-test scores by 2.5% from 15% to 12.5%. For sample 3A, items 3, 7, 8_10, 11_13, 27_29, 36, 40, and 42 were eliminated to produce a fair instrument. By removing these items, the original gender gap in FMCE post-test scores for sample 3A was reduced by 1% from 12% to 11%.

4. Pretest analysis

The above suggests that while some FMCE items perform differently for men and women, most items perform consistently with the overall difference in post-test score. This, however, does not eliminate the possibility of a general bias in the instrument shared approximately equally by all items. Henderson *et al.* [20] explored overall instrumental fairness by binning students by pretest score. Instrumental bias should affect all students regardless of preparation because bias is a property of the test itself, not the student population. As such, any instrumental bias should be observed in all samples and in all bins. If bias is not observed in all bins or in all samples, it would provide evidence that the instrument itself was not biased. Figure 2 plots the male and female FMCE post-test scores binned by FMCE pretest score. The FMCE has items with many more distractors than either the FCI or the CSEM, and therefore this plot is binned somewhat differently than in Henderson *et al.* [20].

To analyze the gender gap in the pretest bins in Fig. 2, linear regression was used to explore the overall gender differences, then t tests with a Bonferroni correction were used to calculate differences in the individual bins. The regression used post-test percentage as the dependent variable and bin number and gender (coded with women as 0 and men as 1) as the independent variables. In Fig. 2(b), too few women scored in the range 13–14 or 15–16 for analysis and, therefore, these bins were eliminated.

For sample 1 linear regression (bin number coded 1–7) yielded a significant main effect of bin [$B = 8.09$, $\mathrm{SE} = 0.44$; $t(1,546) = 18.52$, $p < 0.001$] and a significant main effect of gender [$B = 5.28$, $\mathrm{SE} = 1.28$; $t(1,546) = 4.12$, $p < 0.001$] where B is the regression coefficient and SE is the standard error of the coefficient. The bin-by-gender interaction was not significant. As such, men scored 5.28% higher than women independent of pretest bin. *Post hoc* analysis with a Bonferroni correction showed the difference in post-test performance between men and women was only significant in one of the pretest bins, bin 7-8, with a small effect size [$t(182) = 3.02$, $p < 0.05$, $d = 0.37$].

For sample 3A linear regression yielded a significant main effect of bin [$B = 7.53$, $\mathrm{SE} = 0.82$; $t(3,045) = 9.20$, $p < 0.001$] and a significant main effect of gender [$B = 3.21$, $\mathrm{SE} = 1.25$; $t(3,045) = 2.57$, $p = 0.010$]. The bin-by-gender interaction was also significant

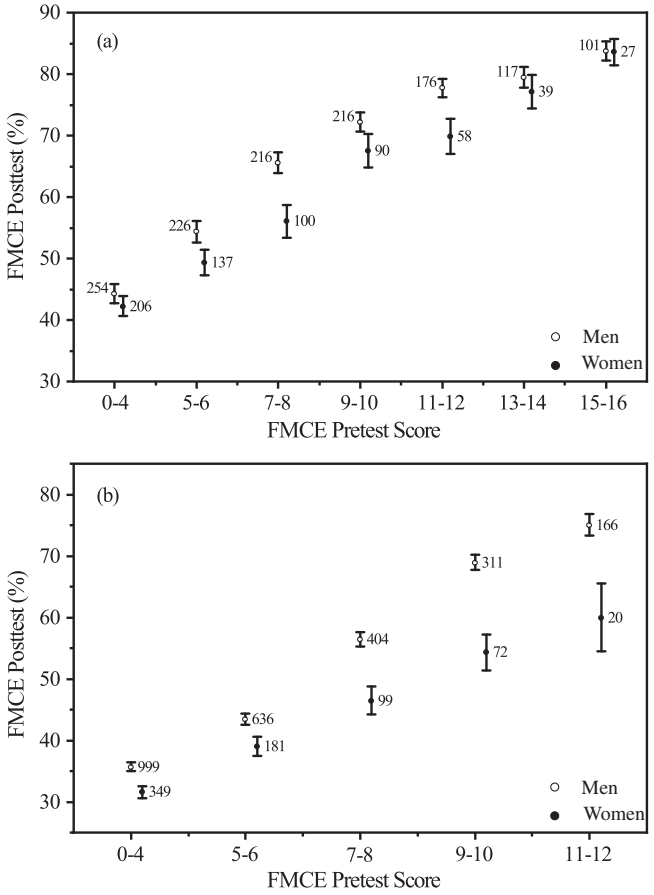

FIG. 2. FMCE post-test percentage vs FMCE pretest: (a) Sample 1 and (b) sample 3A. The number next to the data point is the number of students within each pretest range.

$[B = 3.34,\ \mathrm{SE} = 0.92;\ t(3,045) = 3.62,\ p < 0.001]$. As such, men scored 3.33% higher than women and the difference in score grew with pretest bin number. *Post hoc* analysis with a Bonferroni correction showed the gender gaps in three of the five pretest bins, including bin 0-4, were significant ($p < 0.05$) and the effect sizes grew from a small effect size of $d = 0.20$ in bin 0-4 to a medium effect size of $d = 0.66$ in bin 11-12.

This analysis, which compared men and women with similar pretest scores, showed a general advantage to men of 3%–5% suggesting the FMCE is generally unfair to women by of 3%–5%, a number much smaller than the overall gender gap reported in Table II. This advantage was not equally distributed over all pretest bins in all samples. If the differences measured were the result of instrumental bias, one might expect to observe the same differences in all bins in all samples. The failure to find significant differences in most bins in Sample 1 offers some evidence that the origin of the gender differences by bin might not be instrumental bias; however, this observation can only be viewed as suggestive and more research is needed.

Figure 2 also reports the number of students in each bin; Table V summarizes the percentage of women in each bin for all samples. In all samples, the percentage of women in each bin decreases with increasing pretest score.

TABLE V. The percentage of women in each pretest score bin for each sample. For sample 1, $N = 781$ students had pretest score > 16. For sample 2, $N = 250$ students had pretest score > 12. For sample 3A, $N = 482$ students had pretest score > 12. For sample 3B, $N = 308$ students had pretest score > 12. The number of students in other bins can be found in Figs. 2 and 4.

Sample 1								
Bin	0–4	5–6	7–8	9–10	11–12	13–14	15–16	> 16
% Women	45	38	32	29	25	25	21	14

Sample 2					
Bin	0–6	7–8	9–10	11–12	> 12
% Women	32	29	22	18	8

Sample 3A						
Bin	0–4	5–6	7–8	9–10	11–12	> 12
% Women	26	22	20	19	11	11

Sample 3B					
Bin	0–6	7–8	9–10	11–12	> 12
% Women	25	22	15	17	9

5. Supplemental Material

The results for the mean difficulty P, mean discrimination D, phi coefficient ϕ, and $\Delta\alpha_{\mathrm{MH}}$ for all post-test items of all samples are included in the Supplemental Material [80].

B. CSEM

Sample 2 and sample 3B were analyzed to explore gender differences on the CSEM. Overall averages are presented in Table VI. For sample 2, a gender difference of 4% and 6% was measured on the CSEM pretest and post-test, respectively. These differences in performance were significant: CSEM pretest $[t(1060) = 9.61,\ p < 0.001,\ d = 0.43]$ and CSEM post-test $[t(763) = 6.89,\ p < 0.001,\ d = 0.36]$ with small effect sizes. Results for sample 3B were similar with men outperforming women by 4% on the CSEM pretest $[t(895) = 8.30,\ p < 0.001,\ d = 0.35]$ and by 5% on the

TABLE VI. CSEM pretest and post-test averages for samples 2 and 3B. Averages are reported as percentages.

		Men		Women	
	N	N	$(M \pm \mathrm{SD})\%$	N	$(M \pm \mathrm{SD})\%$
Sample 2					
CSEM Pretest	2108	1618	29 ± 11	490	25 ± 8
CSEM Post-test	2014	1552	65 ± 16	462	59 ± 16
Sample 3B					
CSEM Pretest	3185	2642	27 ± 11	543	24 ± 9
CSEM Post-test	2657	2155	46 ± 18	502	41 ± 17

CSEM post-test [$t(780) = 6.06$, $p < 0.001$, $d = 0.29$] also with small effect sizes.

1. Item analysis

Table VII presents the problematic items for Sample 2 and Sample 3B with item difficulty and item discrimination outside the desired ranges. In Sample 2, all of the problematic pretest items for women had $P < 0.2$ except item 4 with $D < 0.2$, while the majority of the problematic pretest items for men had $P < 0.2$ except for items 21 and 27 which had $D < 0.2$. The results for the Sample 3B pretest were similar; the majority of problematic items had $P < 0.2$ for both men and women, except for item 4 and 21 for women and item 21 for men which had $D < 0.2$. Overall, men and women demonstrated little incoming knowledge of electricity and magnetism in both samples.

Table VII also presents the problematic CSEM post-test items for sample 2 and sample 3B. Post-instruction the number of problematic items was reduced for both men and women in both samples. Although there was very little commonality in the CSEM post-test problematic items between sample 2 and sample 3B, within each of the samples there were many common problematic items between men and women.

In the sample 2 post-test, items 1, 12, 23, and 26 were problematic for both men and women ($P > 0.8$). In addition, for men, items 3, 6, and 19 also had $P > 0.8$. For women, item 31 had $P < 0.2$ and item 32 had $D < 0.2$.

TABLE VII. CTT problematic items with $P < 0.2$, $P > 0.8$, or $D < 0.2$ for the CSEM.

Gender	Pre or post	Problematic items
		Sample 2
Women	Pre	4, 5, 7, 10, 11, 14, 15, 16, 20, 22, 23, 24, 25, 26, 28, 29, 31
	Post	1, 12, 23, 26, 31, 32
Men	Pre	7, 11, 14, 15, 20, 21, 22, 23, 24, 25, 26, 27, 29, 31
	Post	1, 3, 6, 12, 19, 23, 26
Overall	Pre	4, 7, 11, 14, 15, 20, 21, 22, 23, 24, 25, 26, 27, 29, 31
	Post	1, 12, 19, 23, 26, 32
		Sample 3B
Women	Pre	4, 7, 10, 11, 13, 14, 15, 16, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 31
	Post	14, 21, 22, 29, 31
Men	Pre	11, 13, 14, 15, 16, 20, 21, 22, 24, 25, 27, 29, 31
	Post	12, 14, 20, 22, 31
Overall	Pre	7, 11, 13, 14, 15, 16, 20, 21, 22, 24, 25, 27, 29, 31
	Post	14, 22, 29, 31

Most of the problematic CSEM post-test items in sample 2 had $P > 0.8$. Only one item was identified as too difficult for women (item 31) and one item failed to discriminate between women who know the material and those that do not (item 32).

In the sample 3B post-test, items 14, 22, and 31 were problematic for men and women postinstruction. Items 14 and 31 had $P < 0.2$ and item 22 had $D < 0.2$. The other problematic items were less consistent for men and women. For men, item 12 ($P > 0.8$) and item 20 ($P < 0.2$) were problematic. Item 29 ($P < 0.2$) and item 21 and 22 ($D < 0.2$) were problematic for women.

2. Graphical analysis

Figure 3 plots the mean difficulties for the CSEM for men and women. Figures 3(a) and 3(c) show many items with very low pretest scores. These scores were sufficiently low to be consistent with random guessing; it seems likely that many of the pretest items that overlap the fairness line do so because neither men nor women could answer them.

In both CSEM post-test samples [Figs. 3(b) and 3(d)], the majority of the error bars do not overlap the fairness line; most items were significantly more challenging for women. In sample 2, there were two items that fell significantly below the fairness line and were more challenging to men (items 18 and 20); however, in sample 3B, items more challenging for men were closer to the fairness line. For sample 2, a chi-squared test showed the difficulties for items 3, 5, 6, 20, and 29 were significantly different for men and women with small effect sizes measured by the ϕ coefficient. For sample 3B, items 3, 5, 6, 7, 8, 9, 10, 25, and 29 were significantly different, also with small effect sizes.

3. DIF analysis

Table VIII presents the items in the CSEM post-test that have either small to moderate or large DIF. In sample 2, only item 20 demonstrated large DIF (unfair to men), while two other items, 3 and 6, showed small-to-moderate DIF (unfair to women). In sample 3B, only item 32 demonstrated small-to-moderate DIF; this item was moderately unfair to men.

To construct an unbiased instrument, for each sample, items were iteratively removed and DIF was recalculated. Because item 20 was substantially unfair to men in sample 2, removing items 3, 6, and 20 increased the original gender gap by 0.1%. Removing item 32 increased the gender gap in sample 3B by 0.4%.

4. Pretest analysis

Like the FMCE, the above results suggest that the CSEM items are not differentially fair to men and women. Figure 4 presents the CSEM post-test averages as a function of binned CSEM pretest scores for sample 2 and sample 3B. The pretest scores were binned into four ranges, 0-6, 7-8,

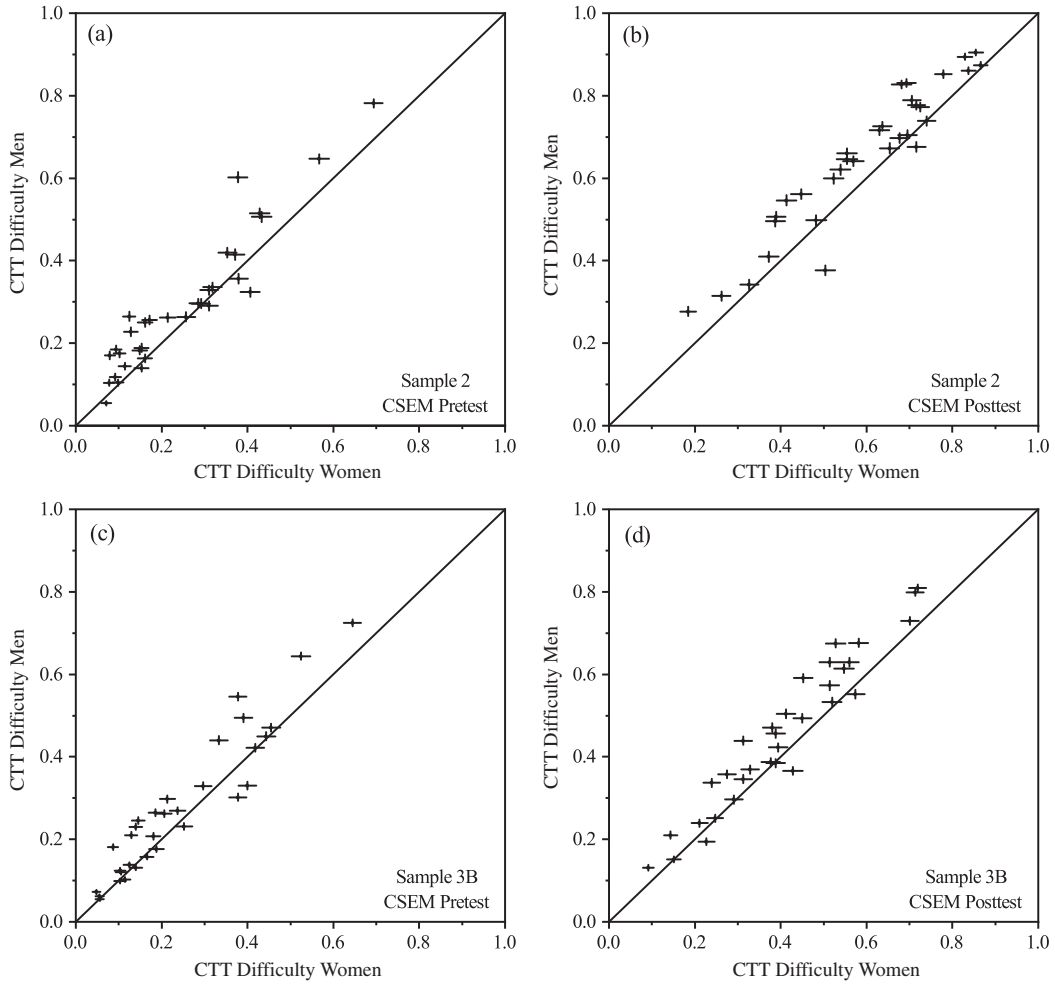

FIG. 3. CTT difficulty results for the CSEM. The top two panels are sample 2 (a) CSEM pretest and (b) CSEM post-test. The bottom two panels are sample 3B (c) CSEM pretest and (d) CSEM post-test. A line of slope 1 is drawn to allow comparison of male and female difficulty. Error bars represent 1 standard deviation in each direction.

9-10, and 11-12; insufficient women scored between 13 and 14 on the CSEM pretest for analysis in both samples.

For sample 2, linear regression with post-test percentage as the dependent variable yielded a significant main effect of bin [$B = 3.70$, SE $= 0.37$; $t(1,550) = 9.90$, $p < 0.001$] and a significant main effect of gender [$B = 3.47$, SE $= 0.88$; $t(1,550) = 3.94$, $p < 0.001$]. The bin-by-gender interaction was not significant. *Post hoc* analysis with a

TABLE VIII. CTT difficulty and discrimination and DIF $\Delta\alpha_{\mathrm{MH}}$ for the CSEM items with small to moderate or large DIF. The significance levels have been Bonferroni corrected: "a" denotes $p < 0.0016$, "b" denotes $p < 0.0003$, and "c" denotes $p < 0.00003$.

Item	P_M	P_F	D_M	D_F	ϕ	$\Delta\alpha_{\mathrm{MH}}$
			Sample 2			
3	0.74 ± 0.01	0.52 ± 0.02	0.67 ± 0.02	0.75 ± 0.03	0.14[c]	-1.01[b]
6	0.83 ± 0.01	0.68 ± 0.02	0.33 ± 0.03	0.55 ± 0.05	0.15[c]	-1.29[b]
20	0.38 ± 0.01	0.50 ± 0.02	0.32 ± 0.03	0.49 ± 0.06	0.11[c]	1.93[c]
Item	P_M	P_F	D_M	D_F	ϕ	$\Delta\alpha_{\mathrm{MH}}$
			Sample 3B			
32	0.37 ± 0.01	0.43 ± 0.02	0.25 ± 0.03	0.33 ± 0.06	0.08[a]	1.02[b]

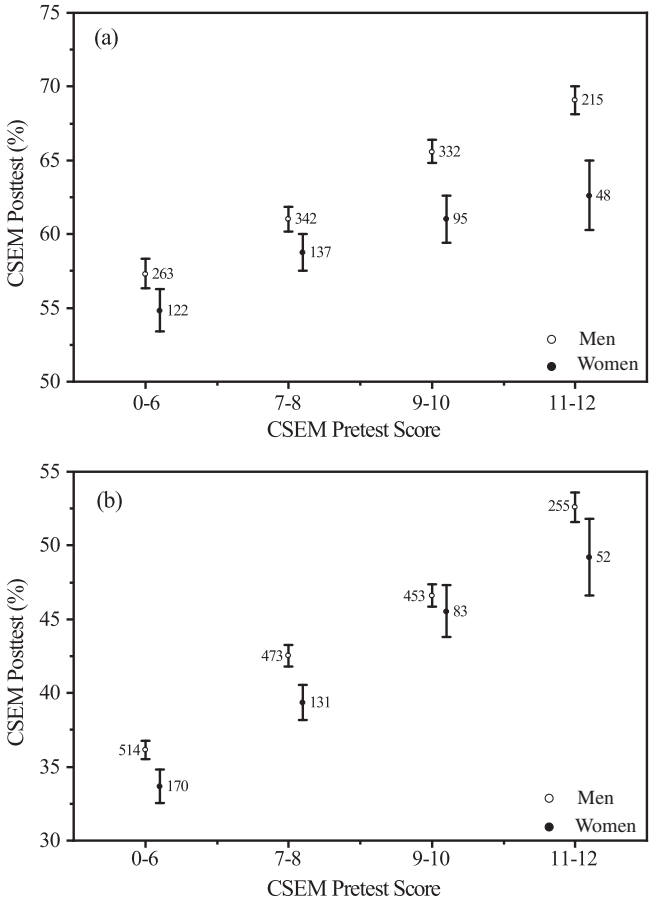

FIG. 4. CSEM post-test percentage vs CSEM pretest: (a) Sample 2 and (b) sample 3B. The number next to the data point is the number of students within each pretest range.

Bonferroni correction showed a significant difference in performance in only bin 9-10 [$t(143) = 2.60$, $p < 0.05$, $d = 0.32$] with a small effect size.

For sample 3B, linear regression yielded a significant main effect of bin [$B = 5.35$, SE $= 0.32$; $t(2, 127) = 16.71$, $p < 0.001$] and a significant main effect of gender [$B = 2.49$, SE $= 0.83$; $t(2, 127) = 2.99$, $p = 0.003$]. The bin-by-gender interaction was not significant. *Post hoc* analysis with a Bonferroni correction showed none of the gender gaps were significant.

As such, men have a 2.5%–3.5% advantage on the CSEM independent of pretest bin. While most bins did not show significant differences, examination of Fig. 4 and the regression results suggests this advantage to men is independent of prior preparation and may represent general instrumental unfairness; however, additional research is need to identify the source of this unfairness.

IV. DISCUSSION

This section will discuss the research questions in the order proposed.

RQ1: Are there items in the FMCE or the CSEM which CTT would identify as problematic? Are the problematic items the same for men and women? Prior to instruction, the majority of the problematic items in the FMCE, including items 36 and 38, were identified as items with difficulty $P < 0.2$; however, items 40 and 43 were identified as easy items ($P > 0.8$) on the FMCE pretest. Overall, the FMCE problematic pretest items were consistent across gender within each of the samples and they were consistent between sample 1 and sample 3A. These findings supported those of Talbot [49] and Ishimoto [50] who both found that items 36 and 38 were too challenging on the pretest.

From the above analysis, which is supported by the work of Talbot [49] and Ishimoto [50], students' understanding of Newton's 3rd law when one object is speeding up (item 36) or slowing down (item 38) is weak prior to physics instruction. For comparison, item 15 on the FCI also addresses the same concept as item 36 on the FMCE [9]. In the study performed by Traxler *et al.*, FCI item 15 was only identified as problematic preinstruction and for only one of the female samples [13].

Problematic items were also identified on the FMCE post-test. The majority of the problematic items in both samples had a difficulty of $P > 0.8$; however, only items 40 and 43 were consistent between men and women in both sample 1 and sample 3A. Although this result also agreed with the work presented by Talbot in which items 40 and 43 remained easy FMCE items after instruction [49], only two out the four items in the *velocity graphs* cluster proposed by Smith and Wittmann were identified as consistently problematic items. In addition to demonstrating $P > 0.8$ on both the FMCE pretest and the FMCE post-test, items 40 and 43 also showed poor discrimination in most student populations. This result shows that students do tend to answer the velocity-time graph items correctly; however, it is difficult to tell if these items are easier because students understand the physical concept or if some other feature of the item is causing students to select the correct response.

The other cluster that was described by Smith and Wittmann (*reversing direction*) [51], which assesses the concept of gravity as a constant downward force, was difficult for women prior to physics instruction for both sample 1 and sample 3A. In sample 3A, two of the three items (8_10 and 11_13) within this cluster remained difficult postinstruction. This result was not consistent across samples. For comparison, item 13 on the FCI also evaluates student understanding of constant downward force of gravity regardless of the motion of the object [9]. Traxler *et al.* identified item 13 as problematic for women prior to physics instruction but not after instruction [13]. In addition, item 27_29, which was also within the *reversing direction* cluster, was not identified as problematic on the FMCE post-test. This item is similar to the other

two items; however, the answers are presented in terms of acceleration rather than in terms of force.

Problematic items were also identified within the CSEM. The results for the CSEM were less consistent between sample 2 and sample 3B than between the two samples for the FMCE. Within each sample, the identified problematic items in the CSEM were fairly consistent between men and women. Prior to any physics instruction, many of the CSEM items were identified to be problematic. Within each sample, many of the problematic items were the same for men and women.

For the CSEM post-test, different problematic items were identified in sample 2 and sample 3B. In sample 2, items 1, 12, 23, and 26 were easy problems for both men and women postinstruction; these items went from being too difficult to too easy for both genders. In sample 3B, items 14, 22, and 31 were identified as problematic for both men and women on the CSEM post-test. Items 14 and 31 were consistently challenging for both genders on the CSEM pretest and the CSEM post-test. Item 22 was a difficult item on the CSEM pretest and had a poor discrimination on the CSEM post-test for both genders.

There were only two items that were problematic on the CSEM post-test across the two samples. Item 12 had a difficulty of $P > 0.8$ for men and item 31 had a difficulty of $P < 0.2$ for women. The inconsistencies in problematic items on the CSEM post-test between the two samples may be due to the large differences in post-test scores between sample 2 and sample 3B (Table VI).

The identification of items with difficulty or discrimination out of some desired range as problematic is conventional in CTT. For both instruments studied many of these problematic items probably resulted from the practice of using the same instrument as both a pretest and a post-test. The identification of some problematic items either preinstruction or postinstruction seems inevitable if a common instrument is to be used at multiple institutions as both a pretest and a post-test; however, the existence of a substantial subset of problematic items in either the pretest or post-test implies these scores should be interpreted with caution. For the CSEM, Henderson *et al.* [20] demonstrated that the low female pretest scores (probably caused by the large number of problematic pretest identified in this study) shifted the female pretest score distribution sufficiently that it substantially overlapped the pure guessing score distribution; and therefore, the pretest scores of women were less predictive of their post-test scores than the post-test scores of men.

RQ2: Are there items in the FMCE or the CSEM which are substantially unfair to men or women? Although the incoming pretest scores were somewhat different for sample 1 and sample 3A, the overall result that the majority of the FMCE items were more difficult for women was consistent between the two samples. Almost all of the items on the FMCE post-test were significantly more difficult for women, but none with more than a small effect size. The only item that was not significantly different in both samples was item 30. This item addresses student understanding of Newton's 3rd law for two objects travelling at the same speed when they collide.

The graphical results for both instruments were quite different than those reported by Traxler *et al.* for the FCI [13]. Graphical analysis identified five substantially unfair items within the FCI post-test; the majority of the FCI items moved toward the fairness line from FCI pretest to FCI post-test. This was not the case for the FMCE; although all of the FMCE items became easier items after instruction (as seen with the overall positive shift in item difficulty), the majority of the FMCE items did not cluster around the fairness line postinstruction in either sample 1 or sample 3A.

The item fairness of the CSEM was also examined graphically. In both samples, students' incoming pretest score was low. Overall, less than half of the items on the CSEM post-test were significantly unfair with one item in each sample (item 20 in sample 2 and item 32 in sample 3B) unfair to men. Overall, the majority of the CSEM items did not demonstrate significantly different difficulty for men and women.

DIF analysis allowed the comparison of item performance under the assumption that the total score on the FMCE was an accurate measure of the conceptual ability. In sample 1, only item 27_29 demonstrated large DIF on the FMCE post-test. In sample 3A, items 27_29 and 40 had large DIF. Item 40 was also identified as problematic because it was too easy; the easiness of the item was not the same for men and women in sample 3A.

The other two clusters that were defined by Thornton *et al.* [29], items 8_10 and 11_13, demonstrated small-to-moderate DIF against women in both samples. Overall, all three of the "all of nothing" clusters, which Smith and Wittmann defined as the reversing direction cluster, showed some gender unfairness toward women.

The number of items that demonstrated large DIF in the FMCE was much smaller than the eight large DIF items initially identified by Traxler *et al.* in the FCI [13]. Overall, the FMCE did not demonstrate the substantial item-level gender unfairness reported for the FCI.

There was also little similarity between the types of items demonstrating large DIF. In the FCI, the five substantially unfair item were item 14 (bowling ball rolling out of airplane, items 21-23 (space shuttle under constant thrust with initial velocity perpendicular to thrust), and item 27 (large box sliding on surface with friction). The FMCE items with large DIF were item 27_29 (samples 1 and 3A) (three questions asking the direction of the acceleration at different points in the trajectory of a coin tossed in the air) and item 40 (sample 3A) (velocity-time graph of a toy car moving at constant velocity).

DIF analysis was also performed for the CSEM. In both sample 2 and sample 3B, only one item, item 20, demonstrated large DIF; this item was biased toward women.

In general, the results for both the FMCE and the CSEM presented in this study were quite different than the results of a substantial set of studies which show item-level unfairness in the FCI [13,18,19]. With this, we concluded that both the FMCE and the CSEM are substantially more gender fair than the FCI at the item level.

RQ3: Are the differences in overall post-test performance between men and women dependent on the student's pretest score?

Linear regression analysis identified a 3%–5% advantage for men on the FMCE post-test controlling for pretest bin (an interaction in sample 3A was measured and, in this sample, the advantage grows with pretest bin). A 2.5%–3.5% advantage toward men was identified in the CSEM. The differences were much smaller than the overall gender differences in the averages of the two instruments. As such, controlling for pretest score, both instruments appear somewhat unfair to women. *Post hoc* analyses showed that the gender differences were not significant in most bins.

If the origin of the differences above was instrumental bias, one would expect differences to be identified in all samples and in all bins. This was not observed. In sample 1, the gender gap for most of the FMCE pretest bins was not significant; both the lowest and highest bins strongly overlap. The failure to find a gender gap in these bins in sample 1 suggests that there is not an overall instrumental bias in the FMCE. In sample 3A, the difference between male and female post-test performance was significant in most bins. Because few pretest bins showed bias in sample 1, it seems likely that the difference in post-test performance in sample 3A was a result of some factor other than instrumental bias. This result can only be viewed as suggestive and more research will be required to determine if the general 3%–5% advantage for men on the FMCE post-test controlling for pretest bin is a result of instrumental bias or some other factor.

Henderson *et al.* [20] presented a similar analysis using a subset of the data in sample 2; the larger data set drawn from the same institution supported their conclusions. The gender gap in the CSEM post-test scores grew as a function of binned pretest score; however, the gender gap in the lowest pretest bins was not significant. In sample 3B, the gender difference was not significant for any pretest bin. While not significant in the individual bins, regression analysis as well as visual inspection of Fig. 4 suggests a small overall advantage toward men. The origin of this advantage may be instrumental bias, but more research is required.

The result that there are no significant gender difference in the lowest CSEM pretest bin is consistent with Kohl and Kuo [36].

DIF analysis cannot eliminate the possibility of overall instrumental unfairness; DIF can only detect differential fairness between items. The gender gaps measured by linear regression analysis are substantially smaller than the overall differences in average observed in each sample. Table V provides a partial explanation; in all samples the percentage of women in each pretest bin decreased with the average pretest score. This overrepresentation of women in the lowest pretest bin has been reported in a number of other studies [20,81,82]. The binning figures and regression analysis indicate generally small differences in post-test performance for equally prepared students; however, in general there is an overall difference in preparation of men and women indicated by the distribution of representation of men and women in the pretest bins. This overall difference in distribution of pretest scores may account for a substantial part of the overall gender gap. More research disentangling the effect of general differences in prior preparation from the effects of instrumental bias is needed.

V. IMPLICATIONS

This work demonstrated that both the FMCE and the CSEM have few items with large DIF while Traxler *et al.* showed that the FCI contains multiple large-DIF items [13]. As such, institutions making decisions on the assessment of instructional practices should consider using the FMCE for mechanics courses and the CSEM for electricity and magnetism courses. Traxler *et al.* constructed a reduced 19-item subset of the FCI which was unbiased and had good reliability metrics; this reduced instrument might also be a good option for assessing mechanics instruction. While the FMCE is a clear choice if one wishes an unmodified published instrument in wide use, the choice between the 19-item FCI and the FMCE is less clear. The FMCE demonstrated relatively large absolute differences measured by the ϕ coefficient particularly in sample 1; many of these differences were larger than those for items detected as large DIF and eliminated from the reduced FCI. The reduced 19-item FCI contains items with substantially smaller ϕ coefficients in Traxler's *et al.* [13] main sample than the FMCE in either sample 1 or sample 3B in this work. While the distribution of pretest scores suggests that the large ϕ coefficients in sample 1 and sample 3A may have resulted from differences in the prior knowledge of men and women, more research is needed to fully understand those differences.

The FCI, FMCE, and CSEM all showed some items that functioned differently for men and women. All three instruments also demonstrated general overall performance differences for men and women; the origin of these general differences is not well understood. As such, it may be inappropriate to use the score on these conceptual instruments to assign course credit.

VI. FUTURE WORK

This work is the third of three papers examining gender differences and fairness in the FCI [13], the FMCE, and the

CSEM [20]. Each work was written to be read independently. Some samples in the papers share instructional environment or student population. Our understanding of the origins of the measured gender differences has also advanced since the writing of the first paper. As such, additional understanding can be developed by synthesizing the three studies. Space considerations prevent us from presenting the synthesis in this work; however, the synthesis is in preparation and will be submitted for publication in the near future. This synthesis may shed additional light on relatively large ϕ values in sample 1 and further inform the choice between the reduced 19-item FCI and the FMCE.

VII. CONCLUSIONS

Traxler *et al.* performed an analysis of the item-level fairness of the FCI for men and women [13]. The current study extended that research to the FMCE and the CSEM. For the FMCE, the majority of the items were significantly more difficult for women both pre- and postinstruction; however, no items stood out as being substantially unfair. There was only one item that demonstrated large DIF in both samples; another item demonstrated large DIF in one sample. Both items were unfair to women. For the CSEM, less than half of the items were of significantly different difficulty for men and women. Only one item in either of the samples demonstrated large DIF; this item was substantially unfair to men. The FMCE and the CSEM contained far fewer large DIF items than the number of large DIF items identified in the FCI by Traxler *et al.* [13]. Regression analysis showed that correcting for pretest score that men had a 3%–5% advantage on the FMCE and a 2.5%–3.5% advantage on the CSEM. DIF analysis examined differences in fairness between items and cannot eliminate the possibility that the origin of the general advantage toward men is that most items are consistently unfair.

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 Grant No. EPS-1003907. We would also like to thank Steven Pollock for his contribution of one of the samples and excellent suggestions on the manuscript.

[1] D. Huffman and P. Heller, What does the Force Concept Inventory actually measure?, Phys. Teach. **33**, 138 (1995).
[2] T. F. Scott, D. Schumayer, and A. R. Gray, Exploratory factor analysis of a Force Concept Inventory data set, Phys. Rev. ST Phys. Educ. Res. **8**, 020105 (2012).
[3] M. R. Semak, R. D. Dietz, R. H. Pearson, and C. W. Willis, Examining evolving performance on the Force Concept Inventory using factor analysis, Phys. Rev. Phys. Educ. Res. **13**, 010103 (2017).
[4] 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).
[5] J. Wang and L. Bao, Analyzing Force Concept Inventory with item response theory, Am. J. Phys. **78**, 1064 (2010).
[6] 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).
[7] T. F. Scott and D. Schumayer, Students' proficiency scores within multitrait item response theory, Phys. Rev. ST Phys. Educ. Res. **11**, 020134 (2015).
[8] E. Brewe, J. Bruun, and I. G. Bearden, Using module analysis for multiple choice responses: A new method applied to Force Concept Inventory data, Phys. Rev. Phys. Educ. Res. **12**, 020131 (2016).
[9] D. Hestenes, M. Wells, and G. Swackhamer, Force Concept Inventory, Phys. Teach. **30**, 141 (1992).
[10] 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).
[11] 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).
[12] N. Jorion, B. D. Gane, K. James, L. Schroeder, L. V. DiBello, and J. W. Pellegrino, An analytic framework for evaluating the validity of concept inventory claims, J. Eng. Educ. **104**, 454 (2015).
[13] 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).
[14] N. J. Dorans, ETS contributions to the quantitative assessment of item, test, and score fairness, ETS Res. Report Series **2013**, 1 (2013).
[15] ETS Standards for Quality and Fairness, Educational Testing Service, Princeton, NJ, https://www.ets.org/s/about/pdf/standards.pdf. Accessed 11/11/2017.
[16] American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, *Standards for Educational and Psychological Testing* (American Educational Research Association, Washington, DC, 2014).
[17] M. Zieky, Fairness review in assessment, in *Handbook of Test Development*, edited by S. M. Downing and T. M.

Haladyna (Lawrence Erlbaum Associates, Hillsdale, NJ, 2006), pp. 359–376.

[18] 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).

[19] 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).

[20] 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).

[21] R. J. Beichner, Testing student interpretation of kinematics graphs, Am. J. Phys. **62**, 750 (1994).

[22] P. V. Engelhardt and R. J. Beichner, Students' understanding of direct current resistive electrical circuits, Am. J. Phys. **72**, 98 (2004).

[23] P. V. Engelhardt, An introduction to classical test theory as applied to conceptual multiple-choice tests, in *Getting Started in PER*, Reviews in PER, Vol. 2, edited by C. Henderson and K. A. Harper (American Association of Physics Teachers, College Park, MD, 2009).

[24] 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).

[25] D. Hestenes and I. Halloun, Interpreting the Force Concept Inventory: A response to March 1995 critique by Huffman and Heller, Phys. Teach. **33**, 502 (1995).

[26] P. Heller and D. Huffman, Interpreting the Force Concept Inventory: A reply to Hestenes and Halloun, Phys. Teach. **33**, 503 (1995).

[27] N. Lasry, S. Rosenfield, H. Dedic, A. Dahan, and O. Reshef, The puzzling reliability of the Force Concept Inventory, Am. J. Phys. **79**, 909 (2011).

[28] C. Henderson, Common concerns about the Force Concept Inventory, Phys. Teach. **40**, 542 (2002).

[29] 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).

[30] S. Ramlo, Validity and reliability of the force and motion conceptual evaluation, Am. J. Phys. **76**, 882 (2008).

[31] 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).

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

[33] 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).

[34] 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).

[35] 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**, 1 (2018).

[36] P. B. Kohl and H. V. Kuo, Introductory physics gender gaps: Pre-and post-studio transition, AIP Conf. Proc. **1179**, 173 (2009).

[37] V. P. Coletta and J. A. Phillips, Interpreting FCI scores: Normalized gain, preinstruction scores, and scientific reasoning ability, Am. J. Phys. **73**, 1172 (2005).

[38] V. P. Coletta, J. A. Phillips, and J. Steinert, FCI normalized gain, scientific reasoning ability, thinking in physics, and gender effects, AIP Conf. Proc. **1413**, 23 (2012).

[39] E. Brewe, V. Sawtelle, L. H. Kramer, G. E. O'Brien, I. Rodriguez, and P. Pamelá, Toward equity through participation in modeling instruction in introductory university physics, Phys. Rev. ST Phys. Educ. Res. **6**, 010106 (2010).

[40] 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).

[41] J. Blue and P. Heller, Using matched samples to look for sex differences, AIP Conf. Proc. **720**, 45 (2004).

[42] T. L. McCaskey, M. H. Dancy, and A. Elby, Effects on assessment caused by splits between belief and understanding, AIP Conf. Proc. **720**, 37 (2004).

[43] T. L. McCaskey and A. Elby, Probing students' epistemologies using split tasks, AIP Conf. Proc. **790**, 57 (2005).

[44] L. McCullough, Gender, context, and physics assessment, J. Int. Womens Studies **5**, 20 (2004); http://vc.bridgew.edu/jiws/vol5/iss4/2.

[45] 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).

[46] 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).

[47] 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).

[48] S. J. Pollock, Comparing student learning with multiple research-based conceptual surveys: CSEM and BEMA, AIP Conf. Proc. **1064**, 171 (2008).

[49] 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).

[50] 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).

[51] 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).

[52] 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).

[53] D. E. Meltzer, Analysis of shifts in students? Reasoning regarding electric field and potential concepts, AIP Conf. Proc. **883**, 177 (2004).

[54] J. Leppävirta, The effect of naïve ideas on students' reasoning about electricity and magnetism, Res. Sci. Educ. **42**, 753 (2012).

[55] Physport, https://www.physport.org. Accessed 8/8/2017.

[56] C. Hieggelke and T. O'Kuma, The impact of physics education research on the teaching of scientists and engineers at two-year colleges, AIP Conf. Proc. **399**, 267 (1997).

[57] The Carnegie Classification of Institutions of Higher Education, Center for Postsecondary Research, Indiana University School of Education, Bloomington, IN, http://carnegieclassifications.iu.edu/. Accessed 9/21/2017.

[58] US News & World Report: Education, US News and World Report, Washington, DC, https://premium.usnews.com/best-colleges. Accessed 4/30/2017.

[59] L. C. McDermott and P. S. Shaffer, *Tutorials in Introductory Physics* (Prentice Hall, Upper Saddle River, NJ, 1998).

[60] 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).

[61] R. K. Thornton, Conceptual dynamics: Following changing student views of force and motion, AIP Conf. Proc. **399**, 241 (1997).

[62] P. Kline, *Handbook of Psychological Testing* (Routledge, New York, NY, 2013).

[63] L. Ding and R. Beichner, Approaches to data analysis of multiple-choice questions, Phys. Rev. ST Phys. Educ. Res. **5**, 020103 (2009).

[64] L. Crocker and J. Algina, *Introduction to Classical and Modern Test Theory* (Holt, Rinehart, and Winston, Mason, OH, 1986).

[65] P. M. Sadler, Psychometric models of student conceptions in science: Reconciling qualitative studies and distractor-driven assessment instruments, J. Res. Sci. Teach. **35**, 265 (1998).

[66] R. S. Lindell, Enhancing College Students' Understanding of Lunar Phases, Ph.D. thesis, University of Nebraska, Lincoln, NE (2001).

[67] W. J. Conover, *Practical Nonparametric Statistics*, 3rd ed. (John Wiley & Sons, New York, NY, 1999).

[68] A. C. Davison and D. V. Hinkley, *Bootstrap Methods and Their Applications* (Cambridge University Press, Cambridge, England, 1997).

[69] P. W. Holland and D. T. Thayer, An alternate definition of the ETS delta scale of item difficulty, ETS Res. Report Series **1985**, 1 (1985).

[70] P. W. Holland and D. T. Thayer, Differential item performance and the Mantel-Haenszel procedure, in *Test validity*, edited by H. Wainer and H. I. Braun (Lawrence Erlbaum Associates, Hillsdale, NJ, 1988), pp. 129–145.

[71] B. E. Clauser and K. M. Mazor, Using statistical procedures to identify differentially functioning test items, Educ. Meas-Issues Pra. **17**, 31 (1998).

[72] R. Zwick and K. Ercikan, Analysis of differential item functioning in the NAEP history assessment, J. Educ. Measure. **26**, 55 (1989).

[73] J. Liu, D. J. Harris, and A. Schmidt, Statistical procedures used in college admissions testing, in *Handbook of Statistics. Vol. 26. Psychometrics*, edited by C. R. Rao and S. Sinharay (Elsevier, Amsterdam, 2007), pp. 1057–1091.

[74] R. Zwick, A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement, ETS Res. Report Series **2012**, 1 (2012).

[75] J. Cohen, A power primer, Psychol. Bull. **112**, 155 (1992).

[76] R. G. Rupert Jr., *Simultaneous Statistical Inference*, 2nd ed. (Springer-Verlag, New York, NY, 2012).

[77] R Core Team, *R: A Language and Environment for Statistical Computing*, (R Foundation for Statistical Computing, Vienna, Austria 2017).

[78] D. Magis, S. Beland, F. Tuerlinckx, and P. De Boeck, A general framework and an R package for the detection of dichotomous differential item functioning, Behav. Res. Meth. Instrum. Comput. **42**, 847 (2010).

[79] A. Canty and B. D. Ripley, boot: Bootstrap R (S-Plus) Functions (2017), R package version 1.3-20.

[80] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevPhysEducRes.14.020103 for male and female students' mean difficulty, mean discrimination, phi coefficient, and $\Delta\alpha_{MH}$ for all post-test items of all samples as well as a plot of male discrimination vs female discrimination for all samples.

[81] L. E. Kost-Smith, S. J. Pollock, and N. D. Finkelstein, Gender disparities in second-semester college physics: The incremental effects of a smog of bias, Phys. Rev. ST Phys. Educ. Res. **6**, 020112 (2010).

[82] S. Bates, R. Donnelly, C. MacPhee, D. Sands, M. Birch, and N. R. Walet, Gender differences in conceptual understanding of Newtonian mechanics: A UK cross-institution comparison, Eur. J. Phys. **34**, 421 (2013).