Exploring the gender gap in the conceptual survey of electricity and magnetism

Rachel Henderson, Gay Stewart, and John Stewart*
West Virginia University, Department of Physics and Astronomy, Morgantown, West Virginia 26506 USA

Lynnette Michaluk
West Virginia University Center for Excellence in STEM Education, Morgantown, West Virginia 26506 USA

Adrienne Traxler
Wright State University, Department of Physics, Dayton, Ohio 45435, USA
(Received 11 May 2017; published 6 September 2017)

The "gender gap" on various physics conceptual evaluations has been extensively studied. Men's average pretest scores on the Force Concept Inventory and Force and Motion Conceptual Evaluation are 13% higher than women's, and post-test scores are on average 12% higher than women's. This study analyzed the gender differences within the Conceptual Survey of Electricity and Magnetism (CSEM) in which the gender gap has been less well studied and is less consistent. In the current study, data collected from 1407 students (77% men, 23% women) in a calculus-based physics course over ten semesters showed that male students outperformed female students on the CSEM pretest (5%) and post-test (6%). Separate analyses were conducted for qualitative and quantitative problems on lab quizzes and course exams and showed that male students outperformed female students by 3% on qualitative quiz and exam problems. Male and female students performed equally on the quantitative course exam problems. The gender gaps within CSEM post-test scores, qualitative lab quiz scores, and qualitative exam scores were insignificant for students with a CSEM pretest score of 25% or less but grew as pretest scores increased. Structural equation modeling demonstrated that a latent variable, called Conceptual Physics Performance/Non-Quantitative (CPP/NonQnt), orthogonal to quantitative test performance was useful in explaining the differences observed in qualitative performance; this variable was most strongly related to CSEM post-test scores. The CPP/NonQnt of male students was 0.44 standard deviations higher than female students. The CSEM pretest measured CPP/NonQnt much less accurately for women ($R^2 = 4\%$) than for men ($R^2 = 17\%$). The failure to detect a gender gap for students scoring 25% or less on the pretest suggests that the CSEM instrument itself is not gender biased. The failure to find a performance difference in quantitative test performance while detecting a gap in qualitative performance suggests the qualitative differences do not result from psychological factors such as science anxiety or stereotype threat.

DOI: 10.1103/PhysRevPhysEducRes.13.020114

I. INTRODUCTION

The difference in the performance of male and female students on many of the conceptual evaluations commonly used in physics education research (PER) is well documented and pervasive. Madsen, McKagan, and Sayre provided an overview and analysis of the "gender gap" [1]. Most research has focused on instruments measuring conceptual knowledge of Newtonian mechanics including the Force Concept Inventory (FCI) [2] and the Force and Motion Conceptual Evaluation (FMCE) [3]. For example, in a large study ($N = 5500$) Docktor and Heller reported that male students outperformed female students by 15% on the FCI pretest and 13% on the post-test [4] even though there was no difference in course grade.

Electricity and magnetism evaluations such as the Conceptual Survey of Electricity and Magnetism (CSEM) [5] and the Brief Electricity and Magnetism Assessment (BEMA) [6] are less well studied. In aggregate, these instruments have demonstrated a gender gap of 3.7% on the pretest and 8.5% on the post-test [1]. The gender gap on these instruments is less consistent with Pollock [7] reporting a negative gender gap. Results from the current study were similar to the majority of electricity and magnetism studies, and showed that women scored 6% lower on average than men on the CSEM posttest, 5% on the pretest.

This research adds to the extensive literature on gender gaps in performance on PER conceptual instruments by

*Corresponding author.
jcstewart1@mail.wvu.edu

providing a study featuring a large sample performed at an institution with a less well academically prepared population than many other studies. It also adds to the literature on gender gaps in electricity and magnetism that have not received the same level of attention as gender gaps in mechanics. This study furthers the understanding of the gender gap by comparing gender gaps observed in the CSEM to student performance on both quantitative and qualitative problems assigned in the course studied. These additional problems were assigned in both higher stakes in-semester examinations and lower stakes quizzes allowing the analysis of the effect of testing conditions on the gender gap.

Section II provides a review of the literature on several possible sources of CSEM gender differences. In Sec. III, the data collected, coding of problems, and classroom context is discussed. In Sec. IV, the overall and disaggregated by pretest score results are presented; structural equation modeling results using the latent variable of Conceptual Physics Performance/Non-Quantitative (CPP/NonQnt) are also presented. In Secs. V–VIII, the results are discussed in light of prior findings, implications for instruction, and future work.

II. BACKGROUND

Factors that might be related to the gender gap include prior preparation in physics, performance on standardized tests, cognitive differences in learning, math and or/science anxiety, and stereotype threat. Additionally, high school course-taking patterns and other sources of conceptual prior knowledge such as informal learning experiences are subject to broad patterns of gender socialization. These patterns have been found to be significant in other male-dominated fields such as computer science [8], and if an informal background is similarly helpful in this area, we might expect to see differences especially on qualitative problems.

This work will treat gender as a binary variable despite the call by Traxler *et al.* for a more nuanced treatment of gender and sex [9]. More complete demographic data was not available to explore additional dimensions of the role of gender in the course studied here.

A. Prior knowledge

Differences in prior preparation in physics between male and female students are well documented. Using data drawn from a nationally representative sample [10], a 2015 National Center for Education Statistics report showed that women enroll in high school physics classes at a lower rate than men, with male students receiving high school physics credit at a 5.6% higher rate than female students [11]. Women take chemistry and advanced biology at significantly higher rates than men. The ACT, the company that administers one of the two major U.S. college entrance examinations, reports ($N = 1\,009\,232$) that in 2016 21% of women and 30% of men met the ACT College Readiness in Science, Technology, Engineering, and Mathematics (STEM) benchmark [12]. Taking physics in high school has been shown to increase physics grades in college [13,14] and, therefore, might improve scores on conceptual evaluations.

Antimirova, Noack, and Milner-Bolotin reported that taking high school physics predicted more variation in FCI pretest score than in FCI post-test score but did not find gender predictive of FCI posttest score [15]. Kost, Pollock, and Finkelstein looked at prior physics knowledge by binning students by their FMCE pretest scores and compared FMCE post-test scores between men and women. They found no difference between men and women in any of the pretest bins [16]. In this work, a "bin" will be defined as range of pretest scores; scores are "binned" if divided into groups by ranges of scores. In contrast, Kohl and Kuo binned on CSEM pretest scores and found gender differences in normalized gain in most of the pretest score bins [17]. Kost-Smith, Pollock, and Finkelstein found that male students outperformed female students by 1.5% on the BEMA pretest, a gap that grew to 6% on the post-test [18].

Kost-Smith, Pollock, and Finkelstein also explored using the FMCE post-test score from the previous class as a measure of prior knowledge. They separated students into five FMCE post-test bins. A higher proportion of women than men were found in the lower FMCE post-test score bins while more men than women were found in the higher bins [18]. Bates *et al.* also found that the lowest performing quartile of students on the FCI pretest consisted of approximately half of the female student population. Most of these female students remained in the lowest performing quartile on the FCI post-test [19].

B. Gender in standardized testing and grades

Gender gaps between male and female student performance on standardized examinations such as the Scholastic Aptitude Test (SAT) or Graduate Record Examination (GRE) have also been documented. The Educational Testing Service's (ETS) Gender Study (1997) provided a nuanced analysis showing gender differences varied by subject and that differences were not uniform within the same subject (male students were better at some mathematics skills, female students at other skills) and that a large gender gap between male and female students that had existed in math and science in 1960 had largely closed by 1990 [20]. The female advantage in language skills had not closed. More recently, the College Board reported that in 2006 male and female students scored approximately equally on the SAT verbal/critical reasoning subtest; however, male students scored 536 on the mathematics subtest while female students scored 502 [21]. This difference represented approximately one-third of a standard deviation. The difference had been approximately constant for the previous decade. The ETS concluded that *"Gender differences are not easily explained by single variables such as course-taking patterns or types of tests. They not only*

occur before course-taking patterns begin to differ and across a wide variety of tests and other measures, but they are also reflected in different interests and out-of-school activities, suggesting a complex story of how gender differences emerge" [20].

The differences observed in standardized test performance are counter to a generally consistent higher performance on course grades by women [20]. Voyer and Voyer provide an overview of this body of research in a meta-analysis of studies involving over one million students at all academic levels K–20 [22]. The female academic advantage was strongest in language classes, Cohen's $d = 0.37$, and weakest in science, $d = 0.15$, and mathematics, $d = 0.07$; however, for classes where female students outnumbered male students, the advantage in math was reduced to $d = 0.03$ and in science $d = 0.01$. The female advantage in mathematics and science grades also became smaller with time from middle school through college. Cohen's effect size conventions are $d = 0.2$ represents a small effect, $d = 0.5$ a medium effect, and 0.8 a large effect [23]. Cohen suggests that results of statistical analyses must be interpreted in terms of practical as well as statistical significance [24]. More recent analysis has suggested that Cohen's original effect size criteria should be adjusted for educational research with medium effects as $d = 0.4$ and large effects as $d = 0.6$ [25].

The gender gap on standardized examinations may be related to the gender gap on conceptual evaluations. Kost, Pollock, and Finkelstein used regression analysis to show that combining the FMCE pretest score along with math placement exam score, Colorado Learning Attitudes about Science Survey [26] pretest, and the semester the physics course was taken, explained 70% of the gender gap in the FMCE post-test [16]. A similar model explained 62% of the gender gap in BEMA post-test scores [18]. Men also outperform women on the FCI post-test when using the SAT math score as a covariate [27]. The gender gap has been shown to be the greatest for students with high reasoning skills (Lawson scores) [28].

C. Cognitive factors

Differences in physics prior knowledge may imply that male students are more likely to be relearning the material than female students. This could have differing effects on the pretest and the post-test. The relation of relearning a complex task to learning it for the first time has been extensively studied [29], was central to the development of early theories of memory [30], and, more recently, has been shown to have a physiological origin [31]. In foundational experiments, Ebbinghaus demonstrated that the more thoroughly a task is initially learned, the more quickly it can be relearned [30]. Patterns of learning and forgetting have also been measured within a physics class, finding substantial fluctuations in student knowledge levels on the same topic within a semester [32].

A large body of literature exists exploring the differences in numerous cognitive abilities between men and women [33]. The evidence for superior male spatial reasoning abilities [34,35] and superior female verbal abilities [36,37] is fairly robust, but these constructs are multidimensional and advantages are not uniform across all subfacets. Conceptual physics problems often involve a mixture of verbal, graphical, and logical reasoning. Cognitive researchers have not yet investigated whether there is a gender-based cognition advantage for either sex in the processes needed to solve conceptual physics problems. Some evidence for a cognitive effect on physics performance has been demonstrated; a program of spatial training was shown to result in improved test performance in introductory mechanics [38]. As such, if cognitive differences are the origin of the gender gap, targeted training may alleviate the differences. Spatial training has proven effective in improving spatial reasoning and shows promise for improving retention of women to STEM [39]. For a review of current research on cognitive sex differences see Miller and Halpern [40].

D. Science and math anxiety

Mathematics anxiety can cause students of both genders to perform more weakly on quantitative assessments. Differences in math anxiety by gender have been investigated [41,42]. The difference in mathematics anxiety between boys and girls had approximately the same effect size, $d = 0.28$, as the difference in mathematics self-efficacy, $d = 0.33$. These differences were substantially larger than the differences in mathematics performance, $d = 0.11$ [42]. Mathematics anxiety has been shown to be negatively correlated with performance, $r = -0.27$ [41], a relation that is independent of gender. The effect size conventions for correlations suggest $r = 0.1$ as a small effect, $r = 0.3$ a medium effect, and $r = 0.5$ a large effect [23].

The phenomenon of science anxiety and its relationship to gender has also been explored [43–46]. Mathematics and science majors have the lowest levels of science anxiety when compared to nonscience majors [47]; however, within these mathematics and science majors, female students were more anxious than male students.

Within the physics classroom, students with more communication apprehension achieved lower gains on the FCI [48]. Physics students that see their instructors as allowing more autonomy had lower anxiety about taking a physics course and demonstrated higher performance [49].

E. Testing conditions

Testing conditions may also influence the gender gap. Conceptual evaluations are often given under low stakes testing conditions where students receive credit for good faith efforts. It is possible that male and female students react differently to testing conditions and that their performance would be changed if the evaluation was given as part of a higher stakes in-semester examination. Significant

differences in exam performance for "low stakes" and "high stakes" applications of the same instrument have been demonstrated with small effect sizes [50]. Higher exam stakes have been shown to be positively correlated with student motivation and performance [51]. The relation of interest and effort on low-stakes science and mathematics test performance has also been demonstrated [51] with interest positively correlated with performance. Unfortunately, these studies have controlled for gender rather than investigated differences by gender. Other testing conditions such as the time limit placed on the examination have not been shown to have a significant effect on performance [20].

F. Stereotype threat

Women are substantially underrepresented in physics [52] and in the engineering disciplines that provide the majority of the enrollment in many calculus-based physics classes [53]. The National Science Foundation reported that in 2014, while women received 57% of all bachelor degrees in the U.S., they received only 19% of those awarded in physics and 20% of those awarded in engineering [54]. As a substantially underrepresented population, the performance of women in physics classes may be influenced by stereotype threat. The effect of stereotype threat on academic performance has been investigated as an explanation of differences in performance of men and women in STEM disciplines [55–57]. Shapiro and Williams define stereotype threat as *"a concern or anxiety that one's performance or actions can be seen through the lens of a negative stereotype—a concern that disrupts and undermines performance in negatively stereotyped domains"* [58]. Studies have shown that stereotype threat does indeed have a negative effect on both women's performance and women's interest in STEM fields [58]. Picho, Rodriguez, and Finnie's meta-analysis examined over 15 years of research, specifically about female performance in mathematics under stereotype threat [59]. The research showed an overall negative effect on the quantitative performance of female students, $d = -0.24$; however, this effect was greater for middle school and high school students compared to college students. Gunderson *et al.* investigated how parents' and teachers' gender-related math attitudes can have a negative effect on women when choosing a STEM or math-related career [60]. Within physics, Koul, Lerdpornkulrat, and Poondej demonstrated a three-way interaction between gender typicality, gender contentedness, and gender stereotypes on physics self-concept [61]. Women who had strong math gender stereotypes and a combination of high gender typicality and gender contentedness had a negative physics self-concept.

G. Instrumental and other effects

Multiple authors have suggested that some items within the FCI [2] exhibit a gender bias [62–65]; however, these results have been inconsistent. The CSEM is substantially less well studied than the FCI and similar studies have not yet been carried out. The item contexts in the CSEM are often fairly abstract (point charges, field maps) unlike the more concrete contexts of the FCI (rockets, planes) and may be less susceptible to gender bias. Differences in the gender gap by item have also been identified in in-semester physics assessments [66] and in problems used in physics competitions [67].

Finally, other factors that may contribute to the gender gap on physics conceptual inventories include method of instruction and the use of a standardized instrument. Multiple studies have shown interactive engagement instructional methods are beneficial in reducing the gender gap on conceptual evaluations [68–70] and improving success in physics classes [71]; however, the reduction of the gender gap has not been replicated in all settings [72]. The use of a standardized instrument may cause mismatches in coverage between the instrument and the class tested, presenting students with problems on which they have received little instruction. This could produce gender differences either through differences in prior knowledge or through differences in the psychological response to being asked to solve problems one should not be expected to answer correctly. The psychological response could interact with stereotype threat.

The results of this study do not advance any claim that gendered patterns reported here are fixed, inherent, or apply equally to any individual student (regardless of their gender identity). In most calculus-based physics courses, 80% or more of the students identify as male. In settings with skewed demographic samples, it is important to ask whether reported learning gains are equally distributed among students, or whether they primarily accrue to students from traditionally overrepresented groups in physics.

H. Research questions

This research seeks to answer the following research questions: *RQ1: Does student performance on the CSEM show evidence of a gender gap in the course studied? RQ2: How does the difference in male and female performance on the CSEM compare with those observed in other problems assigned in the course? Are differences consistent between qualitative and quantitative problems? Are differences consistent between low and high stakes testing conditions? RQ3: Are these differences dependent on the student's CSEM pretest score? RQ4: If a single latent variable is constructed to measure the difference in qualitative and quantitative performance, how does this variable differ by testing conditions? How does this variable differ for male and female students?*

III. METHODS

A. Context for research

The research was conducted in the second-semester, calculus-based physics course at the University of

TABLE I. Class size and gender composition by semester.

Semester	N	Men (%)	Women (%)
Fall 2007	73	78	22
Spring 2008	180	74	26
Fall 2008	71	79	21
Spring 2009	200	75	25
Fall 2009	69	80	20
Spring 2010	179	75	25
Fall 2010	87	78	22
Spring 2011	204	73	27
Fall 2011	117	83	17
Spring 2012	227	81	19

Arkansas, a large midwestern land-grant university serving approximately 25 000 students in the United States. The institution had a Carnegie classification of highest research activity through the period studied. The institution, however, had lower national stature and featured engineering and science graduate programs that ranked lower than those found in many PER studies [73]. At the time of the submission of this work, the undergraduate engineering program was ranked 105th [74]; this ranking was fairly consistent for all semesters studied. Engineering students form the majority of the students (80%) in the class studied. Much of the PER research cited in the introduction was performed at more highly ranked institutions. For example, the University of Colorado-Boulder's undergraduate engineering program was ranked 32nd and Colorado School of Mines 44th at the time the data were accessed [74]. As such, the students studied should be somewhat less academically prepared than those in many previous studies of gender differences in physics. The course studied covered electricity, magnetism, and optics. Most students taking the course were enrolled in engineering or physical science degree programs and elected the course because it was required for their major.

While there was some spring-to-fall fluctuation of overall class size, the gender composition of participating students was fairly consistent for the 10 semesters studied. The class size and the percentage of male and female students is shown for each semester in Table I. Women were substantially underrepresented in the course for all semesters studied.

Students were required to attend two 50-min lectures each week and two 2-h laboratory sessions. Lectures were presented traditionally with attendance managed with an in-class quiz. Homework was due before each lecture session. Homework assignments were divided into an open-response assignment collected on paper before each lecture and a multiple-choice assignment entered electronically before each lecture. Four in-semester examinations and a final examination were used to assess student learning. Laboratory sessions featured a mixture of TA-led demonstrations, small group problem solving, inquiry-based explorations, and traditional laboratories. Students were given a quiz during each laboratory session, a lab quiz, to assess their understanding of the previous homework assignment. The CSEM was used to measure student conceptual understanding gains and was given as a lab quiz pre- and postinstruction; both were graded for credit just as any other lab quiz. All course assignments featured a mixture of conceptual and quantitative problems. The course was presented with few modifications during the period studied. The course was considered effective by the physics department, producing strong learning gains on the CSEM, high course evaluation scores for the lead lecturer and teaching assistants, and encouraged many students to elect physics as a major leading to a strong growth in the number of physics majors graduated [75].

The course studied was designed to be both an excellent learning experience for students and a stable research environment for PER. The same lead instructor presented all lectures, designed all assignments, and oversaw TA training during the time studied. As such, much of the variation present in many courses was minimized.

B. Conceptual survey of electricity and magnetism

This work will compare the pretest and post-test responses on the CSEM to answers to other multiple-choice physics problems in the class studied. The CSEM is a 32-problem multiple-choice instrument containing qualitative problems in electricity and magnetism. The test requires approximately 45 min to administer. Each problem has 5 possible responses. The problems cover a range of electromagnetic topics including the electrostatic behavior of point charges, electric potential, the magnetostatic behavior of electric currents, and induction. The test does not cover Gauss' or Ampere's law, electric circuits, or electromagnetic waves.

C. Identifying nonquantitative problems

Each problem presented in the course was classified as either quantitative or nonquantitative using a rubric developed for a National Science Foundation project (DUE-0535928). This rubric was developed to allow reliable classification while also identifying all problems presented in popular PER conceptual evaluations as nonquantitative. The identification of nonquantitative problems was complicated by the existence of conceptual inventory problems requiring some mathematics (for example, if the distance between two point charges is doubled, how does the electric force change?) or problems that were only superficially quantitative (for example, an object with radius 4 cm and volume charge density 3 $\mu C/m^3$ is stationary at the origin, what is the magnetic field at a point 10 cm along the positive y axis?). The last example contains numbers but requires no calculation and could be converted into a problem that would be identified as quantitative by modifying it to require numeric calculation (for example, an object with radius 4 cm and volume charge density

3 $\mu C/m^3$ is stationary at the origin, what is the electric field at a point 10 cm along the positive y axis?). The rubric was constructed and tested on problems found in popular textbooks. Three raters applied the rubric to problems found in seven textbooks achieving 96% agreement. One rater then used the rubric to classify all problems presented in the course studied.

D. Evaluation environment

The class required students to complete a variety of assignments: homework, quizzes completed in lecture (lecture quizzes), quizzes completed in the laboratory (lab quizzes), and in-semester examinations. Lecture quizzes and homework were often completed cooperatively and, therefore, could not be used as individual measures of understanding. Lab quizzes and in-semester examinations were administered so that each student worked individually. In-semester examinations were composed of both open-response and multiple-choice problems; only the multiple-choice test problems were analyzed in this study. The multiple-choice test problems were fairly evenly divided between qualitative (nonquantitative by the above rubric) and quantitative problems. The average of the qualitative multiple-choice test problems is denoted as test qualitative or "TestQual." The average of the quantitative multiple-choice test problems is denoted as test quantitative or "TestQuant." Lab quizzes were composed primarily of conceptual problems designed to evaluate the students' understanding of the previous homework assignment (not the lab they had just completed). They were taken on computers in the lab room during the lab session. The average of the qualitative lab quiz problems is denoted by lab qualitative or "LabQual." There were insufficient numbers of quantitative lab quiz problems for analysis. The CSEM pretest and post-test were administered and graded as lab quizzes, and therefore, the lab qualitative average measured a second set of qualitative problems given under the same testing conditions as the CSEM.

This study will, then, evaluate the average score for male and female students on five collections of problems: the CSEM pretest, CSEM post-test, qualitative lab quiz problems, qualitative test problems, and quantitative test problems. These problems were administered to students in two testing environments: the lab quiz environment and the in-semester examination environment.

All problems were given postinstruction and were specifically designed for the course (except CSEM problems). As such, all test and lab quiz problems were problems the instructor believed had been covered during the course. The tests formed approximately 70% of the course grade, were administered in large lecture theaters, and were therefore a moderately high pressure experience. Lab quizzes formed only 5% of the course grade, were administered in lab, and were believed to be a much lower pressure experience.

In the class studied, four in-semester examinations were administrated; only the first three are included in this study. The last three weeks of the class and the fourth in-semester examination were devoted to ray optics which is not covered by the CSEM. All ray optics problems were removed from the analysis so that the coverage of the analyzed lab quiz and test problems was the same general coverage as the CSEM. No CSEM problem was used in either the non-CSEM lab quizzes, the in-semester tests, or any other material or assignment in the class.

E. Sample

The data were collected from the Fall 2007 semester to the Spring 2012 semester. During this time, 1851 students completed the class for a grade (77% male and 23% female). Students who did not complete all problems on the CSEM pretest or post-test were eliminated, leaving $N = 1407$ students that formed the sample for the analysis which follows. Multiple-choice responses to all CSEM pretest, post-test, qualitative lab quiz, and test problems were collected from these students which resulted in a data set containing 199 483 responses: CSEM pretest 45 024, CSEM post-test 45 024, qualitative lab quiz 70 749, qualitative test 18 993, and quantitative test 19 693.

F. Bonferroni correction

This work will report multiple statistical tests and as such inflation of the type I error rate should be considered. The large sample size also makes interpretation of significance tests problematic and effect sizes will be reported when possible. A Bonferroni correction adjusts the significance levels for the number of statistical tests by dividing the p value by the number of statistical tests performed. This work will employ 15 statistical tests. A Bonferroni correction would adjust significance levels with $p < 0.05$ becoming $p < 0.0033$, $p < 0.01$ becoming $p < 0.000\,67$, and $p < 0.001$ becoming $p < 0.000\,067$. Few results will be changed by this correction. Most tests produced significance levels of $p < 0.001$; these results were also significant at the $p < 0.000\,067$ level. Uncorrected p values will be reported. Tests that would be modified by the correction will be noted as they are presented in the text. The structural equation modeling analysis and the many statistical tests implied by the analysis were treated as independent and not included in this correction.

IV. RESULTS

Table II summarizes the overall averages separated by gender for each problem collection. On average, male students outperformed female students on each set of qualitative problems including the CSEM pretest (5%), the CSEM post-test (6%), the laboratory quizzes (3%), and the in-semester tests (3%). Male and female students performed equally on in-semester quantitative test problems.

TABLE II. Male and female student averages for different problem collections.

	Male students	Female students
Problem collection	$(M \pm SD)\%$	$(M \pm SD)\%$
CSEM pretest	29 ± 11	24 ± 8
CSEM post-test	66 ± 16	60 ± 16
Lab quiz qualitative	73 ± 12	70 ± 13
Test qualitative	75 ± 16	72 ± 18
Test quantitative	79 ± 16	79 ± 16
N	1084	323

The gender differences were examined using t tests. Significant differences between male and female students were found on the CSEM pretest [$t(729) = 8.59$, $p < 0.001$, $d = 0.46$], the CSEM post-test [$t(531) = 5.92$, $p < 0.001$, $d = 0.37$], qualitative laboratory quiz problems [$t(508) = 3.37$, $p < 0.001$, $d = 0.22$], and qualitative test problems [$t(495) = 2.80$, $p = 0.005$, $d = 0.19$]. Cohen's d was used to characterize the effect size for each collection of problems. Effect sizes ranged from a small effect size for qualitative test average and lab quiz average to small to medium effect sizes for the CSEM pretest and post-test score. There was no significant difference between male and female students on the quantitative test problems. The difference between male and female students on qualitative test problems would not be significant and the difference in the qualitative lab quiz problems would be significant at the $p < 0.05$ level if corrected for the number of statistical tests performed using a Bonferroni correction.

The data set was reduced from the 1851 students who completed the course for a grade to the 1407 student sample for this study by the restriction to students who completed all problems on both the pretest and posttest. If this restriction is relaxed, the pretest and post-test averages change little. For the 1788 students who answered any problem on the pretest, the mean pretest percentage was 27.7% [men 28.6%; women 24.4%], which was very similar to the scores of the 1613 students who answered all pretest problems 27.8% [men 28.9%; women 24.4%]. These values are also very similar to the results for students who answered all pretest and post-test questions in Table II. For the post-test, 1665 students answered any question with an average percentage correct of 64.0% [men 65.2%; women 59.8%]. Of these, 1582 answered all questions with an average percentage of 64.6% [men 65.8%; women 60.3%], also very similar to the paired results in Table II. Blank questions were treated as incorrect in this analysis.

A. The effect of the pretest score

Prior conceptual knowledge was measured by giving the CSEM as a pretest. A density distribution of male and female pretest scores is presented in Fig. 1. Table II and Fig. 1 show that male students have a higher pretest average, but also that the male pretest distribution is skewed with a substantial number of men receiving high pretest scores. The post-test density distribution is plotted in Fig. 2.

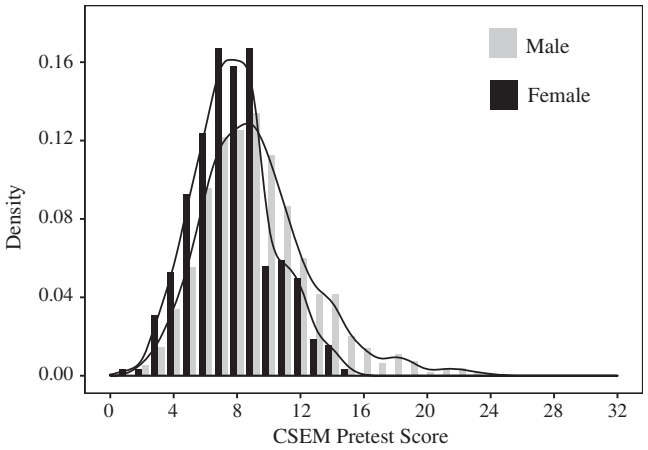

FIG. 1. The distribution of CSEM pretest scores for male and female students. Density plots are drawn for men and women; the male plot reaches its maximum to the right of the maximum of the female plot. The male and female histograms are displaced by the width of 1 bar so that the histograms do not overlap. The density plots are not displaced.

To explore the effects of these differences in pretest scores on students' performance postinstruction, the sample was divided into subgroups. The CSEM is a 32-problem, 5-response evaluation and, therefore, a student should answer 6.4 problems correctly if he or she guesses randomly. To produce groups that contained enough female students for analysis, students were grouped into pretest score ranges (bins) 0–6, 7–8, 9–10, and 11–12. Too few

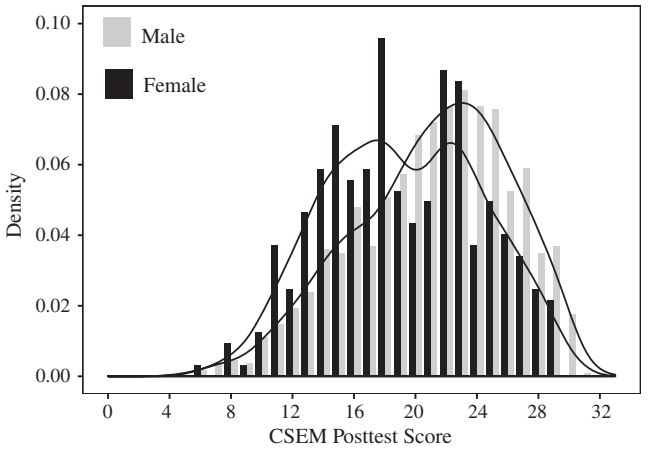

FIG. 2. The distribution of CSEM post-test scores for male and female students. Density plots are drawn for men and women; the male plot reaches its maximum to the right of the maximum of the female plot. The male and female histograms are displaced by the width of one bar so that the histograms do not overlap. The density plots are not displaced.

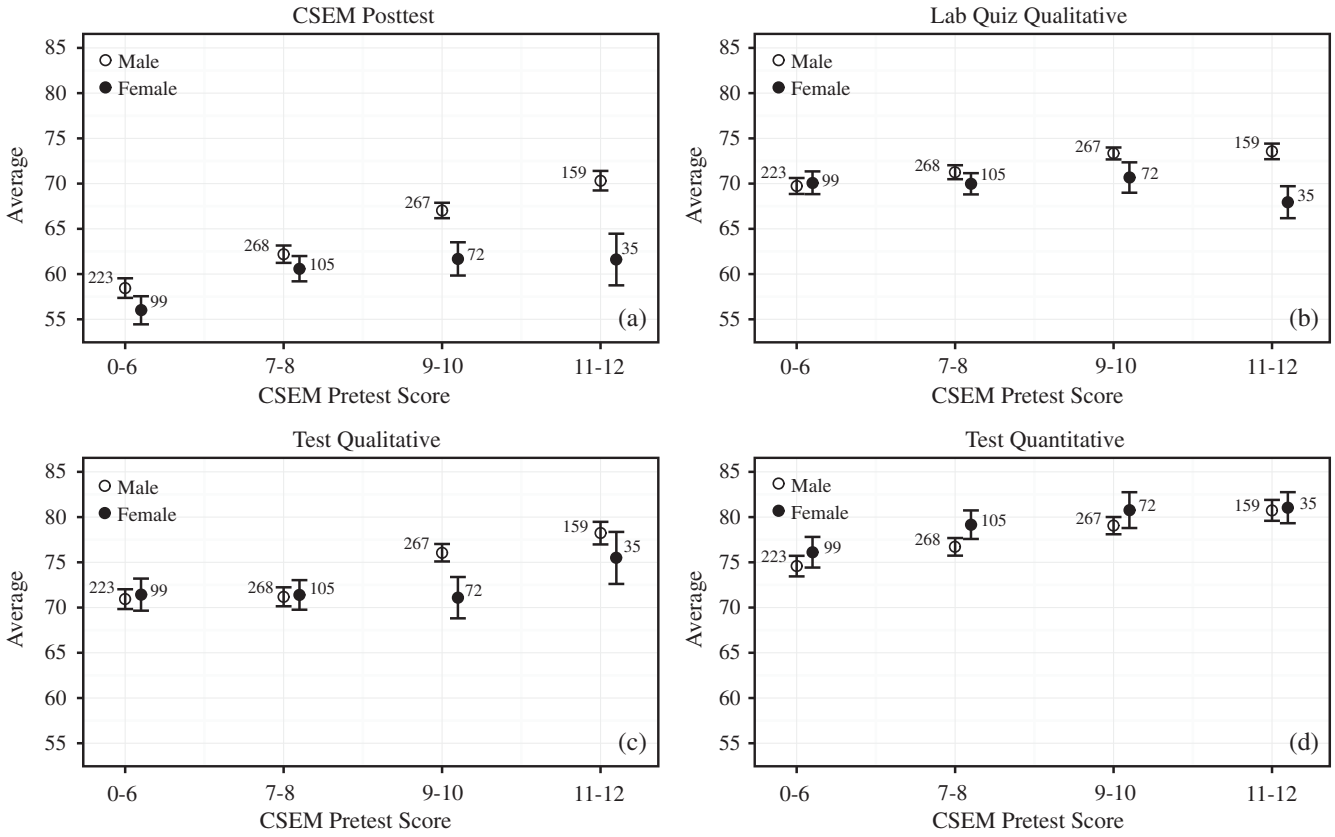

FIG. 3. Evaluation average vs CSEM pretest: (a) the CSEM post-test, (b) qualitative lab quiz problems, (c) qualitative test problems, and (d) the quantitative test problems. Female averages were shifted to the right to increase readability. The number printed next to the point is the number of students within each pretest range.

female students scored 13 or above on the pretest for analysis.

Figure 3 presents the average score within each pretest range for male and female students for each problem collection. For pretest scores between 0 and 8 (bin 0–6 and 7–8), a t test found no significant difference between male and female students in the number of correct responses for any problem collection; therefore, no gender gap exists for pretest scores of 25% or less. Although a small gap of approximately 2% was observed in the CSEM post-test scores for students scoring 25% or less on the pretest, this difference was not significant. The gender gap in the CSEM post-test grew rapidly with the pretest score. A similar, but weaker, relationship between pretest score and gender gap was found in both the qualitative test and lab quiz problem scores. No significant gender gap was found for quantitative test problems; female students outperformed male students particularly at the lowest levels of preparation. The equal quantitative test averages resulted from a greater number of male students with higher levels of preparation who were not plotted in Fig. 3.

B. Latent variable analysis

The qualitative outcomes measured by CSEM post-test score, lab quiz average, and qualitative test average showed similar behavior when plotted against CSEM pretest score, Fig. 3. All have small differences at the lowest pretest score, but a growing difference between male and female outcomes becomes apparent as the pretest score increases. This pattern of increasing gender difference in performance was not observed in the quantitative test results. The similarity of the qualitative results suggested that the difference in qualitative and quantitative performance may be explained by a common latent variable. This variable should be related to the prior conceptual knowledge required for higher pretest scores and any cognitive ability that aids in the solution of qualitative problems but does not contribute to the solution of quantitative problems. As Meltzer noted [76], pretest scores combine prior knowledge with academic ability. We called the latent variable Conceptual Physics Performance/Non-Quantitative or CPP/NonQnt. CPP/NonQnt was functionalized as the part of conceptual performance not explained by overall physics quantitative performance measured by quantitative test average. CPP/NonQnt measures the part of the effect of prior knowledge and conceptual ability that does not result in improved quantitative performance.

Structural equation modeling (SEM) was used to extract CPP/NonQnt and to assess whether it is a productive variable for understanding the differences in conceptual

performance observed. First, to control for general physics ability, the quantitative test average was used as the independent variable in regressions against the qualitative dependent variables: CSEM pretest score, CSEM post-test score, qualitative lab quiz average, and qualitative test average. A latent variable, CPP/NonQnt, was then introduced and used to predict the qualitative variables. CPP/NonQnt was required to be orthogonal to the quantitative test average. The "laavan" package in the R statistical software system was then used to fit the model and the result is shown in Fig. 4. The resulting model had generally good fit parameters. The chi-squared statistic [$\chi^2(2) = 6.21$, $p = 0.045$] was on the border of that required for rejecting the null hypothesis of perfect model fit, and near Kline's $\chi^2/df \leq 3.0$ [77] rule of thumb for good model fit. Weaknesses in the chi-squared statistic, particularly for samples with large N as in this study, causing the incorrect rejection of the null hypothesis of perfect model fit have been extensively researched [77]. Further, the null model for the chi-squared test of perfect model fit is not well aligned with the research question which explores the efficacy of a single latent CPP/NonQnt variable; this assumption is expected to be only approximately true as CPP/NonQnt must certainly be a multidimensional construct.

The weaknesses of the chi-squared test at large N as well as its sensitivity to the features of the underlying distribution and the size of the model correlations [77] have led to a number of additional statistics with superior performance and extensive research into combinations of statistics [78]. This continues to be an active area of research and general rules for SEM fit are still under development [79]. These fit statistics, called approximate fit indices, suggested a good model fit [78]. A wide variety of indices exist; among the most used are the root mean square error of approximation (RMSEA), the standardized root mean square residual (SRMR), and the comparative fit index (CFI). Hu and Bentler [78] found that a combination of two fit statistics dramatically improve the probability of retaining a correct model or rejecting an incorrect model. They suggest RMSEA < 0.05, SRMR < 0.09, and CFI > 0.96 for an acceptable model fit. For the model shown in Fig. 4, RMSEA $= 0.039$, SRMR $= 0.012$, and CFI $= 0.997$; all well within the range of good model fit. The 90% confidence interval of the RMSEA was 0.005 to 0.075. A RMSEA less than 0.05 is considered good fit and greater than 0.10 poor fit; the confidence interval excludes the region of poor fit [78]. All regression coefficients, factor loadings, and variances were significant ($ps < 0.001$). As such, the model fit statistics suggest the latent variable, CPP/NonQnt, produced a model that improves upon a model without the latent variable.

The distribution of male and female CPP/NonQnt is shown in Fig. 5; a density plot of each distribution is also included. The CPP/NonQnt calculated by SEM was normalized by subtracting the mean and dividing by the standard deviation. The difference in CPP/NonQnt between male and female students shown in Fig. 5 was significant [$t(517) = 7.0$, $p < 0.001$, male students $M = 0.10$, $SD = 1.00$, female students $M = -0.34$, $SD = 0.98$]. Because CPP/NonQnt is normalized, differences may be interpreted as Cohen's d effect size and, therefore, the difference between the male and female CPP/NonQnt, 0.44, represents a small to medium effect size.

The binning used in Sec. IV A was repeated in Fig. 6, which demonstrated a growing difference in CPP/NonQnt with the CSEM pretest score, as well as an approximately linear relation between male pretest scores and CPP/NonQnt.

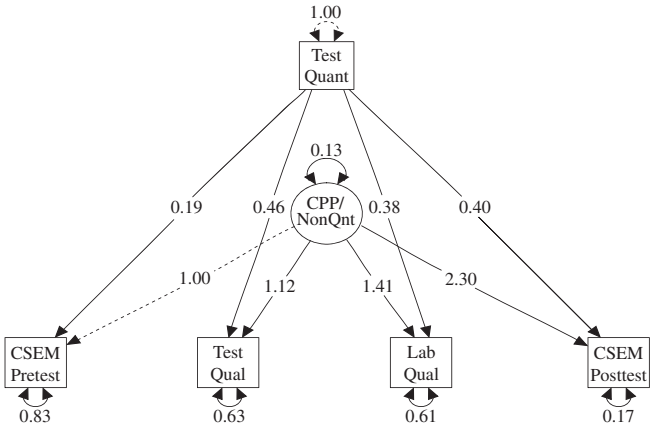

FIG. 4. Structural equation model for CPP/NonQnt's relation to qualitative problem performance. The rectangles represent measured variables and the oval an unmeasured latent variable. The weighting of lines between observed variables are the linear regression coefficients. The weighting of lines between latent and observed variables are the factor loadings. The curved lines represent the variance in each variable.

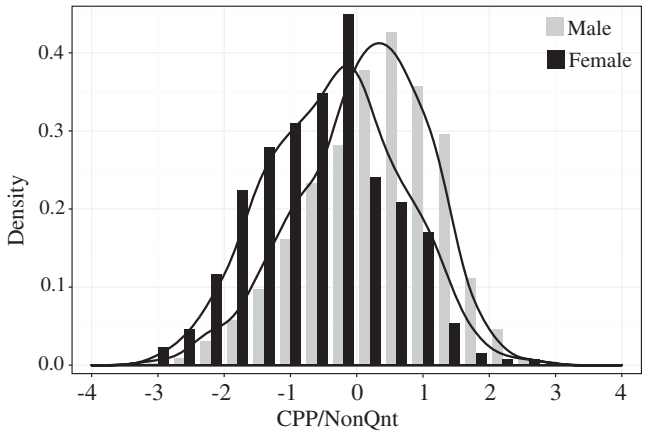

FIG. 5. The distribution of male and female students' CPP/NonQnt. Density plots are drawn for men and women; the male plot reaches its maximum to the right of the maximum of the female plot. The male and female histograms are displaced by the width of 1 bar so that the histograms do not overlap. The density curves were not displaced.

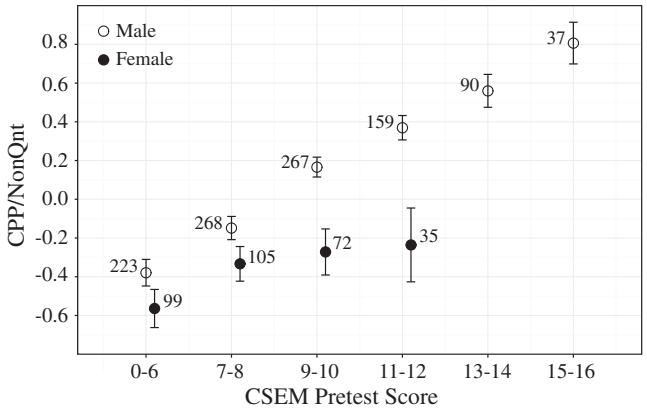

FIG. 6. CPP/NonQnt vs CSEM pretest score.

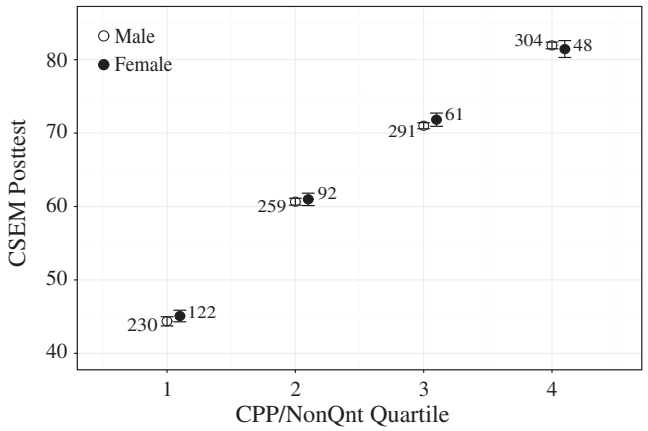

FIG. 7. CPP/NonQnt quartile vs CSEM post-test score.

The relation of pretest score to CPP/NonQnt for women was approximately flat for pretest scores of 7 or more. Correlation analysis was used to explore this qualitative difference. The Pearson correlation coefficient, r, between pretest score and CPP/NonQnt was smaller for women, $r = 0.20$ [$t(321) = 3.57$, $p < 0.001$], than for men, $r = 0.41$ [$t(1082) = 14.86$, $p < 0.001$]. As such, pretest score explained 17% of the variance in CPP/NonQnt for men, but only 4% for women. The correlation between CPP/NonQnt and pretest score for female students would not be significant if corrected for the number of statistical tests performed using a Bonferroni correction.

The differences in CPP/NonQnt were compared for the students with the lowest pretest scores. Combining students with pretest scores of 0 to 8, male and female students had significantly different CPP/NonQnt [$t(400) = 2.4$, $p = 0.018$]; however, this would not be significant if the p threshold was corrected for the number of statistical tests performed with a Bonferroni correction.

While the plots in Figs. 3 and 6 are similar, their interpretation is quite different. Figure 6, and the correlation analysis, suggests that the CSEM pretest scores should be interpreted differently for male and female students with the same pretest score indicating higher CPP/NonQnt for male students.

Figure 7 presents a plot of the CSEM post-test percentage for men and women for each CPP/NonQnt quartile; the quartile was calculated aggregating male and female scores. Male and female students' post-test scores were indistinguishable in each quartile. As such, the growing gender gap observed for all sets of conceptual problems is identified as a result of the differences in the degree to which the CSEM pretest accurately measures CPP/NonQnt for men and women.

If the overall distribution of CPP/NonQnt aggregating male and female students is divided into quartiles, 15% of female students and 28% of male students fall in the highest quartile as shown in Table III. A t test comparing women in the 1st quartile and women in the 2nd and 3rd quartile did not demonstrate a significant difference; therefore, lower and moderately prepared female students are statistically indistinguishable by pretest scores. These students represent 85% of all female participants.

C. Distribution analysis

The observation that pretest scores were more correlated with CPP/NonQnt for male students than female students—that pretest scores measure CPP/NonQnt differently for men and women—warrants further investigation. Figure 1 shows the density distribution of CSEM pretest scores for both male and female students. The pretest scores were very low, and as such, it should be expected that some of the students, who have little knowledge of the material, were guessing. To attempt to understand the differing correlations for men and women, a sequence of models combining binomial distributions representing guessing behavior and normal distributions representing prior knowledge were fit to the distribution of male and female students' pretest scores as shown in Figs. 8(a) and 8(b), respectively.

The dashed lines in Fig. 8 show the result of fitting only a binomial distribution, $B(x; p = 0.2)$, representing pure guessing with probability of success $p = 0.2$ and pretest score x. The pure guessing model was a relatively good fit for female pretest scores. While the fit was not perfect for men, the mean and standard deviation were not that

TABLE III. Male and female student CPP/NonQnt by quartile.

	1st quartile	2nd and 3rd quartile	4th quartile
		Male students	
N	230	550	304
Percentage	21%	51%	28%
$M \pm SD$	7.4 ± 2	9.0 ± 3	11.1 ± 4
		Female students	
N	122	153	48
Percentage	38%	47%	15%
$M \pm SD$	7.3 ± 3	7.7 ± 2	9.0 ± 3

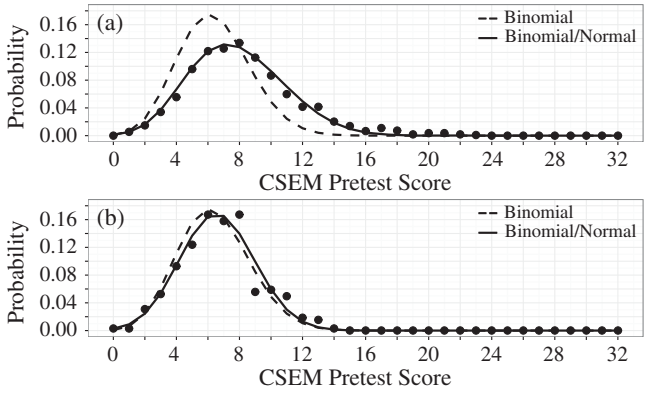

FIG. 8. Model fits for the probability distributions of CSEM pretest scores for (a) male students and (b) female students.

dissimilar from the observed distribution for male students. The solid lines in Fig. 8 plot the result of fitting the model shown in Eq. (1) that mixes a binomial distribution with a normal distribution where p_b is the fraction of students who are guessing, p_n are the fraction of students demonstrating some prior knowledge, and $N(x;\mu_n,\sigma_n)$ is a normal distribution with mean, μ_n, and standard deviation, σ_n:

$$P(x) = p_b B(x; p = 0.2) + p_n N(x;\mu_n,\sigma_n). \quad (1)$$

Fitting Eq. (1) with $p_b + p_n = 1$ yielded $p_b = 0.40$, $p_n = 0.60$, $\mu_n = 8.83$, and $\sigma_n = 2.99$ for the male students. For the female students the fit resulted in $p_b = 0.23$, $p_n = 0.77$, $\mu_n = 6.67$, and $\sigma_n = 2.36$. The curve representing Eq. (1) substantially improves the fit to the male distribution of pretest scores, Fig. 8(a); however, this model did little to improve model fit over the binomial distribution for female students. The mean extracted for the normal distribution for women, 6.67, was very close to the mean of the binomial guessing distribution, 6.40. The difference between the binomial and binomial-to-normal distribution fit for male students suggests that the CSEM can discriminate between male students who exhibit some prior knowledge and those who are guessing. However, for female students the CSEM pretest could not discriminate between those with some prior knowledge and those who were guessing. This analysis explains the qualitative differences in the male and female plots in Fig. 6 and the differences in the correlation of CPP/NonQnt and pretest score. The somewhat lower preparation of women shifts their distribution of pretest scores slightly so that it was less distinguishable from guessing than the male pretest score distribution. As such, the pretest scores of female students provide less information about the incoming knowledge state of the student because of the similarity of pretest results of students with moderate prior knowledge to those with no prior knowledge. This result is almost certainly dependent on the student population; a student body with higher average levels of prior preparation might produce different results.

V. DISCUSSION

This study sought to answer four research questions; these will be addressed in the order proposed.

RQ1: Does student performance on the CSEM show evidence of a gender gap in the course studied? A gender gap of 5% was found in the CSEM pretest and 6% on the post-test. Both these gaps represented small to medium effect sizes. These gaps were consistent with the gaps observed in a large study ($N = 2000$) [18] of the BEMA, but inconsistent with the negative gender gap observed by Pollock ($N = 168$) [7]. The growth of the gender gap from pretest to post-test was consistent with Kohl and Kuo, but of a smaller magnitude [17]. The failure of this study to reproduce the negative gender gap in Pollock could be the result of the less well academically prepared population in this study or differences in instruction.

RQ2: How does the difference in male and female performance on the CSEM compare with those observed in other problems assigned in the course? Are differences consistent between qualitative and quantitative problems? Are differences consistent between low and high stakes testing conditions? Table II shows the gender differences found in CSEM pretest and post-test scores were also present in the other qualitative problems presented in the class; however, the differences were smaller for the other problems (3% for both lab quiz and qualitative test problems). Both these differences represented a small effect size. Male students outperformed female students on qualitative problems in both the low stakes lab quiz environment and the higher stakes in-semester test environment at about equal rates, suggesting that neither the testing rules (low or high stakes) nor the stress of the testing situation were the cause of the gender gap. There was no significant gender gap in the students' quantitative test performance, which provides evidence that the gender gaps observed in the qualitative performance were not the result of general differences in physics ability between male and female students. The CSEM was given in the lab quiz environment, and as such, the larger CSEM post-test gap cannot be attributed to the testing environment.

RQ3: Are these differences dependent on the student's CSEM pretest score? Figure 3 shows that the gender gap was very small at lowest levels of pretest score. No statistically significant difference in CSEM post-test, qualitative lab quiz average, or qualitative test average was found for students with CSEM pretest scores of 25% or less. The gender gap grew with pretest score for all qualitative problem collections. The growth of the gender gap was most pronounced in the CSEM post-test. This result was completely different than that observed by Kost-Smith, Pollock, and Finkelstein [18], where the gender gap disappeared if students were binned by FMCE post-test scores. It was also inconsistent with the CSEM normalized gain results of Kohl and Kuo who found a fairly consistent gender gap, except in the lowest pretest bin [17]. The growth of achievement gaps with

increasing student ability has been well documented [33]; however, the failure to observe any gap in quantitative test scores suggests the growing gender gap observed for qualitative problems had an origin other than in cognitive differences. The students in Kost-Smith, Pollock, and Finkelstein should be substantially more academically prepared than those in this study; in fact, Kost-Smith, Pollock, and Finkelstein [18] report a very small pretest gap. Their failure to observe the growth of the gender gap with pretest score could possibly be explained by a somewhat better prepared female student population which pushed the pretest scores into a range where they were equally predictive of CPP/NonQnt for men and women. The distribution analysis indicates that a small shift in pretest score (Fig. 8) could be enough to greatly change the predictive power of the CSEM pretest.

RQ4: If a single latent variable is constructed to measure the difference in qualitative and quantitative performance, how does this variable differ by testing conditions? How does this variable differ for male and female students? Structural equation modeling demonstrated that a latent variable, CPP/NonQnt, which captured the part of performance on qualitative problems that was not explained by quantitative test average produced a model with good fit. The latent variable had approximately equal effect on qualitative test average, lab quiz average, and CSEM pretest. The variable had a much stronger relation with CSEM post-test scores.

Average male CPP/NonQnt was 0.44 standard deviations higher than female CPP/NonQnt. If the distribution of CPP/NonQnt was divided into quartiles, 13% more male students were in the highest quartile and 17% more female students were in the lowest quartile. This overrepresentation of women in the lowest CPP/NonQnt quartiles is consistent with other research binning students by pretest scores [18,19].

The CSEM pretest score was more weakly correlated with CPP/NonQnt for female students, $r = 0.20$, than for male students, $r = 0.41$. Analysis of the pretest probability distribution suggested that this resulted from the somewhat lower level of female prior knowledge shifting the pretest distribution of moderately prepared women closer to the pure guessing distribution. If CPP/NonQnt rather than pretest score is employed to bin students, no post-test gender gap exists (Fig. 7).

The growing gender gap with pretest score for all qualitative problem collections is well explained by the differential predictive power of CSEM pretest scores for men and women. This also explains the variability in the pretest binning results as the CSEM is applied to academic populations with different levels of preparation. The different correlation of the CSEM pretest scores with CPP/NonQnt for men and women, however, cannot explain the gender differences in the averages of the CSEM pretest, post-test, qualitative lab quizzes, and qualitative tests.

In Sec. I, many potential causes for the gender gap observed in the average scores on conceptual instruments in physics were reviewed. This study was not experimental and cannot conclusively eliminate many of these causes, but a pattern of averages of the different problem collections makes many of these explanations difficult to support. Psychological explanations involving differing responses to testing by gender through math anxiety [41,42], science anxiety [45], or stereotype threat [58] cannot explain why these reactions would occur for qualitative test problems but not quantitative problems on the same test. The failure to find evidence for stereotype threat for this student population further explains the inability to reliably reproduce the effects of interventions to eliminate stereotype threat [80–82] and the failure to detect a relationship between the fraction of women in a class and gender gaps [1]. It seems likely that if efforts to reduce stereotype threat were implemented in the class studied, the gender gap would not be affected.

The observed differences are also difficult to explain by the intrinsic gender fairness of the CSEM instrument. The gender fairness of some FCI items has been questioned [64,65], but no research exists for the CSEM. At pretest scores of 25% or less, no significant gender gap was found. Students who scored less than 25% on the pretest performed more weakly on other class assessments, but the effect was fairly small. It is possible that an intrinsic CSEM gender bias that impacts only the highest performing pretest students exists. This possibility is made less likely by the observation of approximately similar gender gaps in qualitative lab quiz and test scores which did not use CSEM problems.

It is also difficult to resolve the results of this study with an explanation involving cognitive differences between men and women in the ability to solve qualitative physics problems. Cognitive differences vary strongly with the kind of cognitive task [20]. It is possible that men are intrinsically, either through biology or socialization, superior at the combination of verbal, logical, and graphical skills required to solve qualitative physics problems. This explanation seems unlikely; quantitative physics problems like those given in the class studied also require verbal, logical, and graphical reasoning skills, but no gender gap was observed in quantitative problem solving. The quantitative test problems represented a spectrum from problems solvable by substituting numbers into the correct formula to challenging applications of Gauss' law where abstract symbolic and graphical reasoning were required. Further, while male superiority in spatial reasoning [34,35] could impact some qualitative items, one would expect that female superiority at verbal reasoning [36] would be the most important cognitive aspect which differed between qualitative and quantitative problems. As such, one would expect female students to have a cognitive advantage over male students on conceptual problems. No evidence of cognitive abilities differentiated by gender and unique to conceptual physics

problems currently exists; however, research into this aspect of cognition is sparse.

There is at least one explanation for which the observed pattern of averages would be expected. The CSEM pretest is a test of prior knowledge of electricity and magnetism; the problems cannot be answered intuitively without knowing the physical laws. Naturally, a student's academic ability also plays a role, but even a very highly performing student would do poorly on the CSEM if they had no knowledge of the physics tested. The gender gap could be explained by the differences in physics class taking patterns of male and female high school students [10] and differences in informal learning experiences. Both the large CSEM post-test gap and the weaker relation between CPP/NonQnt and qualitative quiz and test averages than with CSEM post-test score could be explained by women overcoming the differences in background while in the class, but men having an advantage on a standardized instrument where coverage was not fully aligned with the class. The large CSEM post-test factor loading in the SEM model could also result if the opportunity to relearn the material instead of learning it for the first time was important in post-test results [30]. Further research should be able to test this conjecture. This interpretation is not fully supported by the work of Kost-Smith, Pollock, and Finkelstein [18] who did not find the years of high school physics taken as a productive variable in predicting post-test scores; however, their analysis used pretest score as an independent variable and, as the authors suggest, high school physics may already have been accounted for in this variable.

Either formal or informal prior physics learning experiences could affect physics performance in many ways. These experiences may produce higher pretest scores, but they may also allow students to master conceptual material more easily by relearning instead of learning for the first time [30]. They may produce higher post-test scores on standardized instruments by filling in holes in coverage. They may also produce more complex interactions such as allowing students retaining misconceptions to confront them again from a different perspective.

This study contributed additional support to previous work showing that mastering quantitative and qualitative problem solving require different learning processes. Students in this sample performed differently on quantitative and qualitative problems given in the same testing environment. The prevalence of poor conceptual performance in noninteractive classes [83] as well as specific experiments investigating the effect of quantitative problem solving on conceptual learning suggest conceptual and quantitative learning are somewhat different processes [84].

VI. IMPLICATIONS FOR INSTRUCTION

The observation that CSEM pretest scores predict CPP/NonQnt and outcomes on qualitative assignments differently for male and female students suggests that pretest scores should be used with caution for instructional decisions such as establishing lab groups or assigning remedial material. The observation that pretest scores are more highly correlated with CPP/NonQnt for men than for women also suggests that the CSEM pretest may be less valid for women than for men [85,86]; that is, a pretest score provides less information about female students than male students. This conclusion is supported by the analysis of the pretest distributions in Sec. IV C.

The persistence of gender gaps for all qualitative problem collections within the course presents a substantial challenge for instruction. Higher levels of CPP/NonQnt benefit students at all points in the course; however, the differences in CPP/NonQnt observed in men and women imply this benefit is not equally distributed for students of different genders. Whether differences in CPP/NonQnt arise from documented differences in high school course taking or less well understood differences in informal education or cognitive processing, women on average have a disadvantage in a physics class when presented with a qualitative problem. CPP/NonQnt loads as strongly on qualitative test average as it loads on pretest score; therefore, differences in CPP/NonQnt have lasting negative effects for women even postinstruction. It is possible that some optional or adaptive remedial strategies could allow women to close the conceptual gap with men. For example, additional qualitative homework problems could be recommended as exam study aids to the entire class. More practice in this area would benefit most students, but could disproportionately help those with lower CPP/NonQnt, which would include many women in this sample but also students who had less high school preparation or less access to informal learning experiences.

The reality is that students in introductory physics courses have extremely variable levels of preparation. The differences identified in CPP/NonQnt between men and women present additional instructional challenges because of a potential interaction between self-efficacy [87] and CPP/NonQnt where male students seem to learn the material more easily because of prior preparation in physics. This could cause women, already with lower self-efficacy toward science [88], to fail to develop self-beliefs consistent with their accomplishments and ability; these women may choose to leave science or engineering careers. This effect has been found in computer science, a field with comparably poor performance in attracting and retaining women [8]. Self-efficacy has been demonstrated to be important in retention [89] and is one of the strongest psychological correlates with academic performance [90]; therefore, it is important as instructional strategies mix students with differing prior knowledge that appropriate support is provided for students who come to the class with less prior knowledge.

VII. LIMITATIONS AND FUTURE DIRECTIONS

This study was performed at a single institution and, therefore, its results may be specific to the student population

or instructional strategy at that institution. The analysis was correlational rather than experimental; additional work is required to understand the relation of CPP/NonQnt to high school preparation, informal learning experiences, and college class taking. Furthermore, additional research is needed to explore whether differences exist in conceptual physics ability differentiated from general physics ability. This study provided evidence that the CSEM as an instrument is not gender biased, but additional item level analysis is needed to determine if the 2% difference in the posttest at lowest levels of preparation results from specific items in the CSEM. While this gap was not significant, the shape of the posttest curves in Fig. 3 and the 2% difference between the posttest and qualitative lab quiz and test averages suggest it warrants further investigation.

The observation that differences in conceptual performance are not related to differences in performance on quantitative problems requires further research. It is unclear if the results of this study would be altered if the pretest and post-test were quantitative and qualitative test performance was used as the control.

The lead instructor of the course was male for all semesters studied. Some research suggests a significant, but weak relationship between the instructor's race or gender and the persistence of students in STEM for students of the same race or gender [91]. Instructor gender effects were also observed in one of the course sections in Kost-Smith Pollock, and Finkelstein's study in which female students outscored male students on participation and homework, but male students scored higher on exams for most semesters studied [18]. In the only lecture section taught by a female instructor, gender differences in exam scores were insignificant. Additional research is needed to determine if the results of the current study would be modified if the lead instructor were female.

VIII. CONCLUSIONS

In this study, gender differences in the CSEM were examined and a 5% gender gap on the pretest was found; the gender gap was 6% on the post-test. This gender gap was also analyzed in other assignments throughout the course: qualitative lab quiz problems, qualitative test problems, and quantitative test problems. The gender gap that was present in the CSEM was also present for the other qualitative problem collections studied. Male students outperformed female students by 3% on both qualitative lab quiz problems and qualitative test problems suggesting that testing environment was not an important source of the gender gap. Male and female students performed equally on quantitative test problems and, therefore, the gender gaps were not a result of general differences in physics ability. The equal performance of men and women on the quantitative test questions also suggests the differences observed in the qualitative questions do not result from psychological factors such as math or science anxiety or stereotype threat. The gender gap for all qualitative problem collections was insignificant for students with a pretest score of 25% or less. The failure to identify a gender gap in either the CSEM pretest or post-test for the least prepared students suggests that there is not an intrinsic gender bias in the CSEM instrument. The gender gap grew with CSEM pretest scores. Structural equation modeling showed that a latent variable called Conceptual Physics Performance/Non-Quantitative, which captured the part of qualitative physics performance not explained by quantitative test average, was productive in explaining the variance in the four qualitative problem sets studied: CSEM pretest, CSEM post-test, lab quiz, and in-semester examination. Male pretest scores were more highly correlated with CPP/NonQnt than female pretest scores and, as such, the pretest is more predictive of CPP/NonQnt for men than for women.

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 through the Taxonomy of Physics Problems (TOPP) Grant No. DUE-0535928.

[1] A. Madsen, S. B. McKagan, and E. C. 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).

[2] D. Hestenes, M. Wells, and G. Swackhamer, Force concept inventory, Phys. Teach. **30**, 141 (1992).

[3] 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).

[4] J. Docktor and K. Heller, Gender differences in both Force Concept Inventory and introductory physics performance, AIP Conf. Proc. **1064**, 15 (2008).

[5] 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).

[6] L. Ding, R. Chabay, B. Sherwood, and R. Beichner, Evaluating an electricity and magnetism assessment tool: Brief Electricity and Magnetism Assessment, Phys. Rev. ST Phys. Educ. Res. **2**, 010105 (2006).
[7] S. J. Pollock, Comparing student learning with multiple research-based conceptual surveys: CSEM and BEMA, AIP Conf. Proc. **1064**, 171 (2008).
[8] J. Margolis and A. Fisher, *Unlocking the Clubhouse: Women in Computing* (MIT Press, Cambridge, MA, 2003).
[9] 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).
[10] C. Nord, S. Roey, S. Perkins, M. Lyons, N. Lemanski, J. Schuknecht, and J. Brown, *American High School Graduates: Results of the 2009 NAEP High School Transcript Study* (National Center for Education Statistics, Washington, DC, 2011).
[11] B. C. Cunningham, K. M. Hoyer, and D. Sparks, *Gender Differences in Science, Technology, Engineering, and Mathematics (STEM) Interest, Credits Earned, and NAEP Performance in the 12th Grade* (National Center for Education Statistics, Washington, DC, 2015).
[12] B. C. Cunningham, K. M. Hoyer, and D. Sparks, *The Condition of STEM 2016* (ACT, Iowa City, IA, 2016).
[13] P. M. Sadler and R. H. Tai, Success in introductory college physics: The role of high school preparation, Sci. Educ. **85**, 111 (2001).
[14] Z. Hazari, R. H. Tai, and P. M. Sadler, Gender differences in introductory university physics performance: The influence of high school physics preparation and affective factors, Sci. Educ. **91**, 847 (2007).
[15] T. Antimirova, A. Noack, and M. Milner-Bolotin, The effect of classroom diversity on conceptual learning in physics, AIP Conf. Proc. **1179**, 77 (2009).
[16] 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).
[17] P. B. Kohl and H. V. Kuo, Introductory physics gender gaps: Pre-and post-studio transition, AIP Conf. Proc. **1179**, 173 (2009).
[18] 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).
[19] 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).
[20] N. S. Cole, *The ETS Gender Study: How Females and Males Perform in Educational Settings* (Educational Testing Service, Princeton, NJ, 1997).
[21] J. L. Kobrin, V. Sathy, and E. J. Shaw, *A Historical View of Subgroup Performance Differences on the SAT Reasoning Test* (The College Board, New York, 2007).
[22] D. Voyer and S. D. Voyer, Gender differences in scholastic achievement: A meta-analysis, Psychol. Bull. **140**, 1174 (2014).
[23] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences* (Academic Press, New York, 1977).
[24] J. Cohen, Things I have learned (so far), Am. Psychol. **45**, 1304 (1990).
[25] J. A. C. Hattie, *Visible Learning: A Synthesis of over 800 Meta-Analyses Relating to Achievement* (Routledge, Taylor & Francis Group, New York, NY, 2009).
[26] 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).
[27] 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).
[28] V. P. Coletta and J. A. Phillips, FCI normalized gain, scientific reasoning ability, thinking in physics, and gender effects, AIP Conf. Proc. **1413**, 23 (2012).
[29] M. Y. Jaber, *Learning Curves: Theory, Models, and Applications* (CRC Press, Tayler & Francis Group, New York, NY, 2016).
[30] A. Baddeley, *Essentials of Human Memory* (Psychology Press, Taylor & Francis Group, New York, NY, 2014).
[31] S. B. Hofer, T. D. Mrsic-Flogel, T. Bonhoeffer, and M. Hübener, Experience leaves a lasting structural trace in cortical circuits, Nature (London) **457**, 313 (2009).
[32] E. C. Sayre and A. F. Heckler, Peaks and decays of student knowledge in an introductory E&M course, Phys. Rev. ST Phys. Educ. Res. **5**, 013101 (2009).
[33] D. F. Halpern, *Sex Differences in Cognitive Abilities*, 4th ed. (Psychology Press, Francis & Tayler Group, New York, NY, 2012).
[34] R. A. Lippa, M. L. Collaer, and M. Peters, Sex differences in mental rotation and line angle judgments are positively associated with gender equality and economic development across 53 nations, Archives of Sexual Behavior **39**, 990 (2010).
[35] Y. Maeda and S. Yoon, A meta-analysis on gender differences in mental rotation ability measured by the Purdue Spatial Visualization Tests: Visualization of Rotations (PSVT: R), Educ. Psychol. Rev. **25**, 69 (2013).
[36] J. S. Hyde and M. C. Linn, Gender differences in verbal ability: A meta-analysis, Psychol. Bull. **104**, 53 (1988).
[37] E. A. Maylor, S. Reimers, J. Choi, M. L. Collaer, M. Peters, and I. Silverman, Gender and sexual orientation differences in cognition across adulthood: Age is kinder to women than to men regardless of sexual orientation, Archives of Sexual Behavior **36**, 235 (2007).
[38] D. I. Miller and D. F. Halpern, Can spatial training improve long-term outcomes for gifted STEM undergraduates?, Learning and individual differences **26**, 141 (2013).
[39] S. A. Sorby, Developing 3D spatial skills for engineering students, Aust. J. Eng. Educ. **13**, 1 (2007).
[40] D. I. Miller and D. F. Halpern, The new science of cognitive sex differences, Trends Cognit. Sci. **18**, 37 (2014).
[41] X. Ma, A meta-analysis of the relationship between anxiety toward mathematics and achievement in mathematics, J. Res. Math. Educ. **30**, 520 (1999).

[42] N. M. Else-Quest, J. S. Hyde, and M. C. Linn, Cross-national patterns of gender differences in mathematics: A meta-analysis, Psychol. Bull. **136**, 103 (2010).
[43] R. A. Alvaro, Ph.D. thesis, Loyola University Chicago, 1978.
[44] J. V. Mallow, A science anxiety program, Am. J. Phys. **46**, 862 (1978).
[45] J. V. Mallow and S. L. Greenburg, Science anxiety: Causes and remedies, J. Coll. Sci. Teach. **11**, 356 (1982).
[46] J. Mallow, H. Kastrup, F. B. Bryant, N. Hislop, R. Shefner, and M. Udo, Science anxiety, science attitudes, and gender: Interviews from a binational study, J. Sci. Educ. Technol. **19**, 356 (2010).
[47] M. K. Udo, G. P. Ramsey, and J. V. Mallow, Science anxiety and gender in students taking general education science courses, J. Sci. Educ. Technol. **13**, 435 (2004).
[48] K. Williams, Understanding communication anxiety and gender in physics, J. Coll. Sci. Teach. **30**, 232 (2000).
[49] N. Hall and D. Webb, Instructor's support of student autonomy in an introductory physics course, Phys. Rev. ST Phys. Educ. Res. **10**, 020116 (2014).
[50] J. S. Cole and S. J. Osterlind, Investigating differences between low-and high-stakes test performance on a general education exam, J. Gen. Educ. **57**, 119 (2008).
[51] J. S. Cole, D. A. Bergin, and T. A. Whittaker, Predicting student achievement for low stakes tests with effort and task value, Contemp. Educ. Psychol. **33**, 609 (2008).
[52] National Science Foundation and National Center for Science and Engineering Statistics, *Science and Engineering Degrees: 1966–2012. Detailed Statistical Tables NSF 15-326* (National Science Foundation, Arlington, VA, 2015).
[53] C. C. de Cohen and N. Deterding, Widening the net: National estimates of gender disparities in engineering, J. Eng. Educ. **98**, 211 (2009).
[54] National Science Foundation, *Women, Minorities, and Persons with Disabilities in Science and Engineering: 2017. Special Report NSF 17-310*, (National Center for Science and Engineering Statistics, Arlington, VA, 2017).
[55] H. D. Nguyen and A. Ryan, Does stereotype threat affect test performance of minorities and women? A meta-analysis of experimental evidence, J. Appl. Psych. **93**, 1314 (2008).
[56] G. Stoet and D. C. Geary, Can stereotype threat explain the gender gap in mathematics performance and achievement?, Rev. Gen. Psychol. **16**, 93 (2012).
[57] G. M. Walton and S. J. Spencer, Latent ability grades and test scores systematically underestimate the intellectual ability of negatively stereotyped students, Psychol. Sci. **20**, 1132 (2009).
[58] J. R. Shapiro and A. M. Williams, The role of stereotype threats in undermining girl's and women's performance and interest in STEM fields, Sex Roles **66**, 175 (2012).
[59] K. Picho, A. Rodriguez, and L. Finnie, Exploring the moderating role of context on the mathematics performance of females under stereotype threat: A meta-analysis, J. Soc. Psychol. **153**, 299 (2013).
[60] E. A. Gunderson, G. Ramirez, S. C. Levine, and S. L. Beilock, The role of parents and teachers in the development of gender-related math attitudes, Sex Roles **66**, 153 2012.
[61] R. Koul, T. Lerdpornkulrat, and C. Poondej, Gender compatibility, math-gender stereotypes, and self-concepts in math and physics, Phys. Rev. Phys. Educ. Res. **12**, 020115 (2016).
[62] L. McCullough, Gender differences in student responses to physics conceptual questions based on question context, ASQ Advancing the STEM Agenda in Education, the Workplace and Society. American Society for Quality, 2011, http://asq.org/edu/2011/06/continuous-improvement/gender-differences-in-student-responses-to-physics-conceptual-questions-based-on-question-content.pdf. Accessed 4/30/2017.
[63] L. McCullough, D. E. Meltzer, M. R. Semak, and C. W. Willis, Differences in male/female response patterns on alternative-format versions of FCI items, *Proceedings of the Physics Education Research Conference 2001, Rochester, NY,* edited by K. Cummings, S. Franklin, and J. Marx (AIP, New York, 2001), pp. 103–106.
[64] 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).
[65] S. Osborne 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, *2011 American Educational Research Association Conference* (American Education Research Association, Washington, DC, 2011).
[66] D. J. Low and K. F. Wilson, Persistent gender gaps in first-year physics assessment questions, *Proceedings of The Australian Conference on Science and Mathematics Education 2015* (The Institute for Innovation in Science & Mathematics Education, Sydney, Australia, 2015), pp. 118–124.
[67] K. Wilson, D. Low, M. Verdon, and A. Verdon, Differences in gender performance on competitive physics selection tests, Phys. Rev. Phys. Educ. Res. **12**, 020111 (2016).
[68] R. J. Beichner and J. M. Saul, Introduction to the SCALE-UP (Student-Centered Activities for Large Enrollment Undergraduate Programs) project, *Invention and impact: Building excellence in undergraduate Science, Technology, Engineering and Mathematics(STEM) education* (American Association for the Advancement of Science, Washington, DC, 2003), pp. 61–66.
[69] M. Lorenzo, C. H. Crouch, and E. Mazur, Reducing the gender gap in the physics classroom, Am. J. Phys. **74**, 118 (2006).
[70] E. Mazur, *Peer Instruction: A User's Manual* (Prentice Hall, Englewood Cliffs, NJ, 1997).
[71] S. W. Brahmia, Improving learning for underrepresented groups in physics for engineering majors, AIP Conf. Proc. **1064**, 7 (2008).
[72] 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).
[73] US News & World Report: Education Best Graduate Schools Physics, https://www.usnews.com/best-graduate-schools/top-science-schools/physics-rankings. Accessed 4/30/2017.

[74] US News & World Report: Education Best Undergraduate Engineering Programs, https://www.usnews.com/best-colleges/rankings/engineering. Accessed 7/5/2017.

[75] J. Stewart, W. Oliver III, and G. Stewart, Revitalizing an undergraduate physics program: A case study of the University of Arkansas, Am. J. Phys. **81**, 943 (2013).

[76] 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).

[77] R. B. Kline, *Principle and Practice of Structural Equation Modeling*, 3rd ed. (Guilford Press, New York, NY, 2011).

[78] L. Hu and P. M. Bentler, Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives, Struct. Eq. Modeling **6**, 1 (1999).

[79] H. W. Marsh, K. Hau, and Z. Wen, In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings, Struct. Eq. Modeling **11**, 320 (2004).

[80] A. Miyake, L. E. Kost-Smith, N. D. Finkelstein, S. J. Pollock, G. L. Cohen, and T. A. Ito, Reducing the gender achievement gap in college science: A classroom study of values affirmation, Science **330**, 1234 (2010).

[81] L. E. Kost-Smith, S. J. Pollock, N. D. Finkelstein, G. L. Cohen, T. A. Ito, and A. Miyake, Replicating a self-affirmation intervention to address gender differences: Successes and challenges, AIP Conf. Proc. **1413**, 231 (2012).

[82] S. Lauer, J. Momsen, E. Offerdahl, M. Kryjevskaia, W. Christensen, and L. Montplaisir, Stereotyped: Investigating gender in introductory science courses, CBE Life Sci. Educ. **12**, 30 (2013).

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

[84] E. Kim and S. Pak, Students do not overcome conceptual difficulties after solving 1000 traditional problems, Am. J. Phys. **70**, 759 (2002).

[85] L. Crocker and J. Algina, *Introduction to Classical and Modern Test Theory* (Holt, Rinehart and Winston, New York, 1986).

[86] 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).

[87] B. J. Zimmerman, Self-efficacy: An essential motive to learn, Contemp. Educ. Psychol. **25**, 82 (2000).

[88] L. M. Larson, K. M. Pesch, S. Surapaneni, V. S. Bonitz, T. F. Wu, and J. D. Werbel, Predicting graduation: The role of mathematics/science self-efficacy, J. Career Assess. **23**, 399 (2015).

[89] R. W. Lent, S. D. Brown, and G. Hackett, Toward a unifying social cognitive theory of career and academic interest, choice, and performance, J. Vocat. Behav. **45**, 79 (1994).

[90] M. Richardson, C. Abraham, and R. Bond, Psychological correlates of university students' academic performance: a systematic review and meta-analysis, Psychol. Bull. **138**, 353 (2012).

[91] J. Price, The effect of instructor race and gender on student persistence in STEM fields, Econ. Educ. Rev. **29**, 901 (2010).