2018

An Exploration of Diversity and Inclusion in Introductory Physics

Rachel J. Henderson

An Exploration of Diversity and Inclusion in Introductory Physics

Rachel J. Henderson

Dissertation submitted
to the Eberly College of Arts and Sciences
at West Virginia University

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in
Physics

John Stewart, Ph.D., Chair
Gay Stewart, Ph.D.
Paul Cassak, Ph.D.
Chandralekha Singh, Ph.D.
Paul Hernandez, Ph.D.

Department of Physics and Astronomy

Morgantown, West Virginia
2018

Keywords: physics, education, gender, conceptual inventories

ABSTRACT

An Exploration of Diversity and Inclusion in Introductory Physics

Rachel J. Henderson

Diversity and inclusion has been a concern for the physics community for nearly 50 years. Despite significant efforts including the American Physical Society (APS) Conferences for Undergraduate Women in Physics (CUWiP) and the APS Bridge Program, women, African Americans, and Hispanics continue to be substantially underrepresented in the physics profession. Similar efforts within the field of engineering, whose students make up the majority of students in the introductory calculus-based physics courses, have also met with limited success. With the introduction of research-based instruments such as the Force Concept Inventory (FCI), the Force and Motion Conceptual Evaluation (FMCE), and the Conceptual Survey of Electricity and Magnetism (CSEM), differences in performance by gender began to be reported. Researchers have yet to come to an agreement as to why these "gender gaps" exist in the conceptual inventories that are widely used in physics education research and/or how to reduce the gaps.

The "gender gap" has been extensively studied; on average, for the mechanics conceptual inventories, male students outperform female students by 13% on the pretest and by 12% post instruction. While much of the gender gap research has been geared toward the mechanics conceptual inventories, there have been few studies exploring the gender gap in the electricity and magnetism conceptual inventories. Overall, male students outperform female students by 3.7% on the pretest and 8.5% on the post-test; however, these studies have much more variation including one study showing female students outperforming male students on the CSEM.

Many factors have been proposed that may influence the gender gap, from differences in background and preparation to various psychological and sociocultural effects. A parallel but largely disconnected set of research has identified gender biased questions within the FCI. This research has produced sporadic results and has only been performed on the FCI. The work performed in this manuscript will seek to synthesize these strands and use large datasets and deep demographic data to understand the persistent differences in male and female performance.

Dedicated to my family.
For their love and endless support.

ACKNOWLEDGMENTS

This dissertation would not have been possible without the endless support and encouragement from my advisor, Dr. John Stewart. Over the years, his motivation and positive feedback enabled me to successfully complete this work. All of the valuable discussions and lessons learned during this experience will certainly be something I will always cherish throughout my career. John, thank you for your guidance throughout this process and, most importantly, for your dedication and commitment to me and my career. I am truly grateful to have you as an advisor, a mentor and a friend.

I would also like to thank my dissertation committee members, Dr. Gay Stewart, Dr. Paul Cassak, Dr. Chandralekha Singh, and Dr. Paul Hernandez for their extraordinary feedback on this manuscript. A special thanks to Gay for always finding the time out of your extremely busy schedule to give me career advice and motherly words of wisdom; your kindness and support over the last four years will forever be cherished.

In addition, thank you to the rest of the physics education research group at WVU: Dr. Paul Miller, Dr. Seth Devore, Dr. Lynnette Michaluk, and fellow office mate and friend, Cabot Zabriskie. Paul, thank you for your insight and guidance into PER; without you, these opportunities may have never happened. Seth, thank you for all of your hard work and collaboration in the various research projects we did together. Lynn, thank you for your endless amounts of literature and kind words throughout the writing process. And finally, a special thanks to Cabot, for always making me laugh even in the most stressful of times. Mostly random, but sometimes research, conversations over lunch and coffee breaks allowed us to build a lasting friendship that I will value forever.

Lastly, I owe my sincerest of thank yous to my parents, Keith and Cindy Henderson. There is no amount of words that I could write to tell you how much I appreciate everything you have done for me. Without your guidance and encouragement, none of this would have been possible. Thank you for all of the sacrifices you have made to make sure I was molded into the goal-driven, strong and ambitious woman I am today. I will forever be grateful for the infinite amount of love and support you both have given to me.

Contents

List of Tables

List of Figures

Chapter 1

Introduction to Physics Education Research

Physics Education Research (PER), which began in the 1980s, is a relatively new field of study. Over the course of approximately 50 years, discipline-based educational research (DBER) has made many strides toward understanding the teaching and learning of physics. A recent synthesis divided the majority of the research done in physics education into six topical areas: conceptual understanding, problem solving, curriculum and instruction, assessment, cognitive psychology, and attitudes and beliefs about learning and teaching [1]. This chapter will review the history of PER, specifically the events leading up to the development of research-based materials, as well as discuss some of the most commonly used materials within the physics education community.

1.1 Prior to the 1980s

Prior to the 1980s, there was much debate on what influenced student success in a physics course. Mostly, researchers agreed that mathematical skills are correlated with success with physics content [2–5]. After determining that students could only answer 80% of basic math questions correctly, Larkin and Brackett saw the need to develop a math-review unit. Students, prior to entering the physics course, received materials with the essential mathematics topics needed to be successful in the course [2]. Hudson and McIntire provided evidence that success in a physics course is influenced by certain math skills [3].

Although there was significant agreement about the correlation between basic math skills and success in the physics classroom, there was substantial disagreement about the role of logical reasoning in physics success [4, 6]. Cohen, Hillman and Agne presented a study of 195 students at the University of Vermont that showed that the Piagetian level of four specific tasks (Floating, Pendulum, Shadows, and Chemicals [7]) was weakly correlated with student success [6]; however, Liberman and Hudson found that both students' logical test scores and students' SAT scores were correlated with the final examination scores [4].

1.2 Student Conceptual Understanding of Physics

In the late 1970s, researchers were also beginning to realize that students develop preconceived notions about the physical world around them [8–10]. In 1980, Champagne, Klopfer, and Anderson conducted a study on the factors that influenced the learning of classical mechanics, which included variables such as these preconceptions, logical reasoning, and mathematical skill [5]. The authors developed an assessment, called the Demonstration, Observation, and Explanation of Motion test (D.O.E), to evaluate the student's preconceptions about motion. This, along with logical reasoning and mathematical skills, were then used to study how these variables influenced students' mastery of classical mechanics. In agreement with prior studies, math skills were highly correlated with achievement in classical mechanics; however, in addition to math skills, both logical reasoning scores and students' D.O.E. test scores demonstrated significant correlations with success in classical mechanics [5]. This study led physics educators to begin to not only focus on students' prior math skills but to also emphasize the importance of students' conceptual understanding of physical concepts [11–16].

In 1982, John Clement performed a qualitative study showing that many students have an alternate view of the relationship between force and acceleration [17]. Through three examples of common physical systems (a swinging pendulum, a coin toss, and a moving rocket), Clement showed that students believe that “motion implies a force.” This misconception was also shown to be extremely stable; post-instruction results showed that many students were still providing incorrect explanations for the three systems. Clement suggested “further development of innovative instruction techniques that emphasize rigorous understanding of qualitative principles” [17]. Following this study, McDermott suggested that physics instruction needed to explicitly address students’ difficulty with conceptual understanding [18].

The PER community had a “breakthrough” in 1985; Halloun and Hestenes constructed a “mechanics diagnostic test” to explore the initial knowledge state of college physics students [19]. This instrument, designed to assess students’ qualitative conceptions of physics, consisted of basic kinematic and dynamic conceptual questions. An analysis of the scores, in relation to student performance in physics, showed that the initial beliefs students have about the world around them have a large effect on their success in the classroom. This idea of developing an assessment to evaluate student conceptual understanding of physics was the beginning of research-based materials.

1.2.1 Physics Conceptual Inventories

Because student success in the classroom was related to conceptual understanding of physical systems and students came into the classroom with these misconceptions about the physical world, the PER community began to develop research-based assessments in an attempt to measure the change in students’ conceptual understanding of physics, specifically, how instruction affects student’s alternative conceptions of the world around them.

In 1985, Halloun and Hestenes conducted a study where they interviewed students about their views of motion while taking the “mechanics diagnostic test” [19, 20]. The interviews consisted of conversations about general concepts of motion, free particle motion, and one- and two-dimensional motion under a constant force. After analyzing student comments, the authors organized students' common sense concepts about motion into two general categories: principles of motion and influences on motion.

In 1992, the first physics conceptual inventory, the Force Concept Inventory (FCI), was developed by Hestenes, Wells, and Swackhamer using the qualitative information on the common sense beliefs that students have about physical systems [21]. Although the FCI is the most commonly used physics conceptual inventory, other assessments have been developed to try to gain insight into students' alternate conceptions of physics. The most popular of these assessments are the Force and Motion Conceptual Evaluation (FMCE) [22], the Conceptual Survey of Electricity and Magnetism (CSEM) [23], and the Brief Electricity and Magnetism Assessment (BEMA) [24].

Force Concept Inventory

The FCI is a 30-item assessment which measures conceptual understanding of one- and two-dimensional kinematics, Newton's laws, and the understanding of forces [21]. Each item has five possible responses and incorrect responses were constructed to match commonly held misconceptions. The FCI was revised after its initial publication; this thesis work will use the revised instrument published with Mazur [25] which is available at PhysPort [26].

Force and Motion Conceptual Evaluation

The FMCE is a 43-item conceptual inventory evaluating students' conceptual understanding of Newton's Laws of Motion [22]. 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. Each block of items includes at least seven 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 [26].

Conceptual Survey of Electricity and Magnetism

The CSEM is a 32-item conceptual inventory evaluating students' conceptual understanding of electricity and magnetism [23]. 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 [27]. 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, vector addition of electric forces, electric field, and magnetic field, as well as induction. The CSEM does not cover Gauss' or Ampere's law, electric circuits, or electromagnetic waves. For a list of all of the concepts evaluated by the CSEM, see Maloney *et al.* [23]. The final version of the CSEM is also available at PhysPort [26].

1.2.2 Formalized Research-Based Materials

In 1992, research-based materials in PER were beginning to become formalized. McDermott and Shaffer wrote a two-part investigation that supported the use of research as a guide for curriculum development in the physics classroom [28, 29]. In agreement with prior studies [17], Part I of McDermott and Shaffer's investigation found that students' conceptual difficulties with basic electrical concepts remained after a "standard presentation of material in the tradition lecture and laboratory format" [28]. The authors suggested that it is extremely valuable for students to actively engage in the learning process in order to overcome their conceptual difficulties; a hands-on experiment in the laboratory is simply not enough. Part II discussed the process which the authors took in designing and developing specific instructional materials to address student conceptual difficulties [29]. In this work, Shaffer and McDermott used an electric circuit module (previously developed by the University of Washington Physics Education Group) to (1) understand student thinking and difficulties with circuits and (2) develop, through many iterations, an instructional material to guide students through their conceptual understanding of circuits. The authors called for this process to be performed for all topics that are covered in a standard introductory physics course.

In 1997, Hake quantitatively investigated the use of interactive engagement on student conceptual understanding [30]. From student pre- and post-instruction FCI scores, Hake defined a measure of course effectiveness: normalized gain. Normalized gain $\langle g \rangle$ is defined as the ratio of the average gain from pre- to post-instruction to the maximum average gain

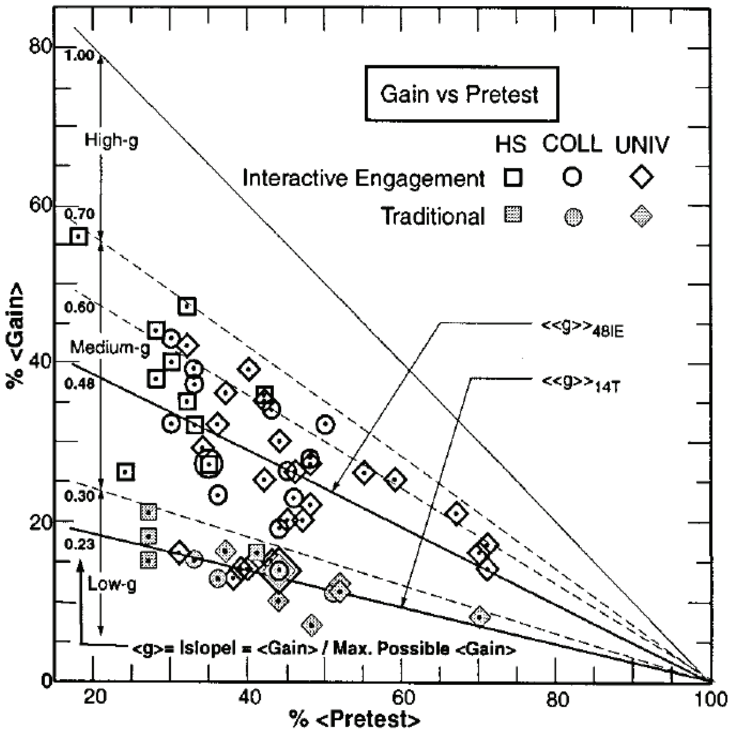

Figure 1.1: Gain vs. pretest for traditional instruction and interactive engagement [30].

possible,

$$\langle g \rangle = \frac{\langle S_f \rangle - \langle S_i \rangle}{1 - \langle S_i \rangle}, \tag{1.1}$$

where $\langle S_i \rangle$ is the class pretest average and $\langle S_f \rangle$ is the class post-test average. Hake classified high-g as $\langle g \rangle \geq 0.7$, medium-$g$ as $0.7 > \langle g \rangle \geq 0.3$, and low-$g$ as $\langle g \rangle < 0.3$.

Hake analyzed the average normalized gains in 64 introductory physics courses, 14 of which were "traditional" courses and 48 of which applied the use of "interactive-engagement" methods. Figure 1.1 plots gain vs. pretest score; the slope is the normalized gain and shows that traditional methods, in general, produce smaller gains than interactive methods [30]. Figure 1.1 quantitatively supports the research program proposed by McDermott and Shaffer [28, 29]; the use of interactive-engagement methods in the physics classroom helps student conceptual understanding of physics.

1.3 Common PER Research-Based Resources

Since the seminal research performed by McDermott and Shaffer [28], Shaffer and McDermott [29] and Hake [30], the PER community has been developing and implementing various teaching methods in the attempt to positively impact student success in the physics classroom. This section will provide an introduction to some of the commonly used PER research-based materials and instructional methods used to increase students' conceptual understanding. The resources discussed below are ones that are, or have been, employed in the Department of Physics & Astronomy at West Virginia University. Additional resources, including expert recommendations along with other teaching methods and assessments are available on PhysPort [26].

1.3.1 Tutorials

Tutorials are the most well-known research-based materials used in the PER community. Following the call from McDermott and Shaffer suggesting a systematic process for the development of research-based materials for *all* areas of physics, the Physics Education Group at the University of Washington developed the *Tutorials in Introductory Physics* [31]. Based on more than 20 years of research, these materials were designed as guided-inquiry worksheets for small groups in introductory calculus-based physics. They "contain questions that try to break the reasoning process into steps of just the right size for students to stay actively involved" [32]. The *Tutorials in Introductory Physics* were written to be a supplement to lectures, labs, and textbooks with teaching assistants guiding students through the worksheets using the Socratic method to help students build their conceptual understanding of physics. More recently, the University of Maryland has developed open source tutorials in an attempt to make it easier for physics educators to adapt this resource to their instructional environment [33–35].

1.3.2 Learning Assistant Model

The University of Colorado at Boulder developed a program in which Learning Assistants (LAs), typically undergraduate students, assist faculty members in the teaching of physics in large-enrollment introductory physics courses [36]. The LAs are trained through a pedagogy course to lead discussions between students to encourage interactive engagement in the physics classroom [37]. The goals of this program are the following: "(1) to improve the education of all science and mathematics students through transformed undergraduate education and improved K-12 teacher education, (2) to recruit more future science and math teachers, (3) to engage science faculty more in the preparation of future teachers and discipline-based educational research, and (4) to transform science departmental cultures to value research-based teaching as a legitimate activity for professors and our students" [36]. Overall, the model has been successful in both improving student conceptual knowledge of physics along with increasing a student's physics identity [38, 39]. The model is also thought to aid in recruiting future physics teachers.

1.3.3 Interactive Lecture Demonstrations and Clickers

Interactive Lecture Demonstrations (ILDs) and Peer Instruction [25, 40] using clickers are tools to get students talking about physics in a large lecture class. The goal of this practice is first to elicit students' misconceptions then to allow them to talk through the difficulties they have about a physical concept in a collaborative manner. With both resources, prior to either a demonstration or a written physics question, students are asked to individually predict the result followed by discussing their predictions with their peers in small groups. Since students commit to a prediction, the demonstration or the solution to a physics problem can help them confront their misconceptions. In the end, either the demonstration is carried out or the answer to the question is given (or provided by the class) and, collectively, the class compares their predictions with the outcome. There are many variations which may involve students reporting their selected answers and reasoning, re-voting, or re-voting after some instructor intervention. In practice, both of these research-based resources could be used as part of the process. Many studies have found that ILDs and clickers are excellent ways to improve the learning of fundamental physics concepts [41–45].

1.3.4 Physics and Everyday Thinking

Physics and Everyday Thinking (PET) is a curriculum designed for a one-semester, conceptual-based physics course, generally targeted at future elementary teachers [46]. Students enrolled in this course are guided through a deep understanding of conceptual physics topics and are expected to engage in small group discussions, whole-class discussions, and hand-on laboratories. Many resources, including an instructor resource DVD, teacher guides, and introductory workshops, are available to assist instructors with implementing this curriculum in their classroom [47]. The development of PET is ongoing but has been shown to improve student success, both in understanding the conceptual content as well as improving their attitudes about science [48].

Overall, the implementation of research-based methods in the physics classroom has been shown to improve student success [38–45, 48, 49]; however, research has also been performed to investigate the barriers to the use of research-based instructional strategies [50]. Sustaining these educational reforms within a physics department with a variety of physics educators can be more challenging than one may expect [51].

PER has many other areas of emphasis including problem solving [2, 52], cognitive psychology [53] and attitudes and beliefs about teaching and learning [54, 55]. These topical areas, along with conceptual understanding and assessment, have also been investigated in both precollege physics students [56, 57] and upper-division physics students [58–60]. While this chapter introduced the relevant history of the conceptual inventories and other research-based resources needed for the research studies in this manuscript, there exists an extensive body of research in many other topics in PER. These topics will not be discussed in detail; however, the following section will discuss the role of gender in relation to the differences in introductory physics.

Chapter 2

The Gender Gap*

This chapter will review the relevant literature of performance differences in physics between male and female students. All of the research studies treat gender as a binary variable and while this treatment may obscure the complicated nature of gender identity [62], the following research in this manuscript will also treat gender as a binary variable, as is common practice in PER studies.

*This summary is drawn from “Exploring the gender gap in the Conceptual Survey of Electricity and Magnetism” published in Physical Review Physics Education Research [61]. This work was constructed with collaborative efforts from Adrienne Traxler, Lynnette Michaluk, and Gay Stewart.

The difference in the performance of male and female students on many of the conceptual evaluations commonly used in PER is well-documented and pervasive. Madsen, McKagan, and Sayre provided an overview and analysis of the "gender gap" [63]. Most of the research has focused on instruments measuring conceptual knowledge of Newtonian mechanics including the FCI [21] and the FMCE [22]. On average, male students outperform female students on the mechanics conceptual inventories by 13% on the pretest and by 12% on the post-test. For example, in a large study ($N = 5,500$) Docktor and Heller reported that male students outperformed female students by 15% on the FCI pretest and 13% on the post-test [64] even though there was no difference in course grade.

Electricity and magnetism evaluations such as the CSEM [23] and the BEMA [24] 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 [63]. The gender gap on these instruments is less consistent with Pollock [51] reporting a negative gender gap.

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 [65]. The following sections will explore possible explanations advanced to explain the consistent gender gap seen in conceptual physics performance.

2.1 Prior Knowledge

Differences in prior preparation in physics between male and female students are well documented. Using data drawn from a nationally representative sample [66], 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 [67]. Women take chemistry and advanced biology at significantly higher rates than men. The ACT, the company that administers one of the two major US 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 [68]. Taking physics in high school has been shown to increase physics grades in college [69, 70] 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 post-test score [71]. 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. A "bin" is defined as range of pretest scores; scores are "binned" if divided into groups by ranges of scores. They found no difference between men and women in any of the pretest bins [72]. In contrast, Kohl and Kuo binned on CSEM pretest scores and found gender differences in normalized gain in most of the pretest score bins [73]. 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 [74].

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 [74]. 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 [75].

2.2 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 [76]. 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 sub-test; however, male students averaged 536 on the Mathematics sub-test while female students averaged 502 [77]. This difference represented approximately one-third of one standard deviation. The difference had been approximately constant for the previous decade. The ETS concluded that *"Gender differences are not eas-*

ily 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" [76].

The differences observed in standardized test performance are counter to a generally consistent higher performance on course grades by women [76]. 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 [78]. The female academic advantage was strongest in language classes and weakest in science and mathematics; however, for classes where female students outnumbered male students, the advantage in math and science was reduced. The female advantage in mathematics and science grades also became smaller with time from middle school through college.

The gender gap on standardized examinations may be related to the gender gap on conceptual evaluations. Kost, Pollock, and Finkelstein used an analysis to show that combining the FMCE pretest score along with a math placement exam score, Colorado Learning Attitudes about Science Survey [79] pretest and the semester the physics course was taken, explained 70% of the gender gap in the FMCE post-test [72]. A similar analysis explained 62% of the gender gap in BEMA post-test scores [74]. Men also outperform women on the FCI post-test when using SAT math score as a covariate [80]. The gender gap has been shown to be the greatest for students with high reasoning skills (Lawson scores) [81].

2.3 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 [82], was central to the development of early theories of memory [83], and more recently has been shown to have a physiological origin [84]. In foundational experiments, Ebbinghaus demonstrated that the more thoroughly a task is initially learned, the more quickly it can be relearned [83]. 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 [85].

A large body of literature exists exploring the differences in numerous cognitive abilities between men and women [86]. The evidence for superior male spatial reasoning abilities [87, 88] and superior female verbal abilities [89, 90] is fairly robust, but these constructs are multi-dimensional and advantages are not uniform across all sub-facets. 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 [91]. Miller and Halpern recently published a review of current research on cognitive sex differences [92].

2.4 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 [93, 94]. The difference in mathematics anxiety between boys and girls had approximately the same effect as the difference in mathematics self-efficacy. These differences were substantially larger than the differences in mathematics performance [94]. Mathematics anxiety has been shown to have a negative relationship with performance [93], a relationship that is independent of gender.

The phenomenon of science anxiety and its relationship to gender has also been explored [95–98]. Mathematics and science majors have lower levels of science anxiety when compared to non-science majors [99]; 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 [100]. Physics students that see their instructors as allowing more autonomy had lower anxiety about taking a physics course and demonstrated higher performance [101].

2.5 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 [102]. Higher exam stakes have been shown to be positively correlated with student motivation and performance [102]. The relation of interest and effort on low-stakes science and mathematics test performance has also been demonstrated [103] 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 [76].

2.6 Stereotype Threat

Women are substantially underrepresented in physics [104] and in the engineering disciplines that provide the majority of the enrollment in many calculus-based physics classes [105]. The National Science Foundation reported that in 2014, while women received 57% of all bachelor degrees in the US, they received only 19% of those awarded in physics and 20% of those awarded in engineering [106]. 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 [107–109]. 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"* [110]. Studies have shown that stereotype threat does indeed have a negative effect on both women's performance and women's interest in STEM fields [110]. Picho, Rodriguez and Finnie's meta-analysis examined over 15 years of research, specifically about female performance in mathematics under stereotype threat [111]. The research showed an overall negative effect on the quantitative performance of female students; 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 [112]. Within physics, Koul, Lerdpornkulrat, and Poondej demonstrated a three-way interaction between gender typicality, gender contentedness, and gender stereotypes on physics self-concept [113]. Women who had strong math gender stereotypes and a combination of high gender typicality and gender contentedness had a negative physics self-concept. Maries, Karim, and Singh investigated stereotype threat in relation to male and female performance on the FCI and the CSEM [114]. The authors found that asking the students to report their gender prior to taking the FCI and the CSEM did not impact their overall performance on the conceptual inventories.

2.7 Instrumental and Other Effects

Multiple authors have suggested that some items within the FCI [21] exhibit a gender bias [115–118]; however, these results have been inconsistent. The FMCE and the CSEM are 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. The item contexts in the FMCE are equally as concrete as the FCI. Differences in the gender gap by item have also been identified in in-semester physics assessments [119] and in problems used in physics competitions [120].

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 [25, 121, 122] and improving success in physics classes [123]; however, the reduction of the gender gap has not been replicated in all settings [73, 124–126]. 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.

Overall, there exists a large number of research studies that explore the differences in physics performance between male and female students; however, physics education researchers have yet to come to an agreement on why a gender gap exists on the commonly used conceptual inventories or how to reduce this gap. The research presented in this manuscript will demonstrate features of the conceptual inventories that are not equally fair for male and female students and suggest ways to remove this unfairness.

Chapter 3

Statistical Methods

Prior to presenting analyses and results, this chapter will summarize the quantitative statistical methods used throughout this manuscript.

3.1 Descriptive Statistics

Descriptive statistics are typically the first step to analyzing data and are used to quantify what is being measured in the dataset. More often than not, descriptive statistics are summarized as a table, graph, or a single-value statistic, such as an average or standard deviation. The following sections will describe the descriptive statistics used in this manuscript.

3.1.1 Central Tendency

Measures of central tendency are used to characterize the most representative observations in a dataset. Mean, median, and mode are common measures of central tendency. A population mean μ is defined as the sum of a set of observations divided by the total number of observations. The median is defined as the middle value of an ordered list of data. The mode is defined as the observation that appears most often.

When describing data, scientists often use the sample mean M to summarize the data. The sample mean has three characteristics: (1) the sample mean is an unbiased estimator of the population mean μ, (2) a distribution of sample means obeys the central limit theorem and (3) a distribution of sample means has minimum variance [127].

3.1.2 Variability

Variability is characterized as how observations vary from the mean. Variance σ^2 and standard deviation σ, which is the square root of the variance, are common measures of variability. Sample variance s^2 is defined as the sum of the squared deviations of observations from the sample mean divided by the degrees of freedom ($df = n - 1$, where n is the number of observations), which makes it an unbiased estimator.

3.2 Inferential Statistics

Inferential statistics are statistical techniques used to predict population outcomes based upon the results from sample data. Hypothesis testing, also known as significance testing, is a method used to determine the likelihood that an observation happened by chance. In behavioral research, the level of significance is set to 5% for significant outcomes; an outcome is significant if there is a 5% probability or less that the observed outcome happened by chance. A significance level less than 5% ($p < 0.05$) is considered statistically significant and should be investigated further [128]. The following sections in this chapter will describe techniques that use inferential statistics to analyze data.

3.2.1 *t*-tests

A basic inferential statistic used to compare means between two groups is called a t-test [129]. This statistic can be used for either two independent groups (e.g., male and female students) or when an observation is repeated twice for the same subject (repeated-measures design). The t-statistic for a two-independent-sample t-test is

$$t = \frac{M_1 - M_2}{s_{M_1 - M_2}}, \tag{3.1}$$

where M_1 is the sample mean of the first independent group, M_2 is the sample mean of the second independent group, and $s_{M_1 - M_2}$ is the estimated standard error of the difference.

This standard error is defined as

$$s_{M_1-M_2} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}, \tag{3.2}$$

where

$$s_p^2 = \frac{s_1^2(df_1) + s_2^2(df_2)}{df_1 + df_2} \tag{3.3}$$

is the pooled sample variance with sample variance s_1^2 of the first independent group, sample variance s_2^2 of the second independent group, and df_1 and df_2 are the degrees of freedom of the first and second independent groups, respectively. For equal sample sizes, the pooled sample variance simplifies to $s_p^2 = \frac{s_1^2+s_2^2}{2}$.

To determine if the two independent sample means are statistically different, the calculated t-statistic is compared to the critical values of the t-distribution [129] with $df = df_1+df_2$ degrees of freedom. If the difference is reported as statistically significant ($p < 0.05$), then the difference between the two independent sample means is taken to be a real effect; there is a small probability the difference happened by chance.

The t-statistic for a repeated-measures design is similar to the two-independent-sample t-test; however, to eliminate between-person error, the difference between the paired observations is calculated prior to calculating the t-statistic.

Prior to analyzing a t-test, four assumptions are made: (1) the data is normally distributed, (2) the data was taken from a random sample, (3) each observation is independent from one another, and (4) the variance of the two samples is equal [127].

3.2.2 Effect Size

Significance testing is a process of showing whether or not an effect exists; however, it does not describe the magnitude of the effect. Along with a significant outcome, an effect size is typically calculated and reported. A common effect size that is reported with a significant t-test in educational literature is Cohen's d [130]. Cohen's d is defined as

$$d = \frac{M_1 - M_2}{\sqrt{s_p^2}}, \tag{3.4}$$

where s_p^2 is defined in Eqn. 3.3. Cohen classified a small effect as $d = 0.2$, a medium effect as $d = 0.5$, and a large effect as $d = 0.8$ [130]. Cohen suggested that results of statistical analyses must be interpreted in terms of practical as well as statistical significance [131]. 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 $d = 0.6$ [132]; however, the research presented in this manuscript will report effect size in terms of Cohen's original criteria.

3.2.3 Error and Power

Using a significance level of 5% leaves the possibility of error in the decisions researchers make about an observed outcome. Error is categorized in two types: Type I error and Type II error. Table 3.1 summarizes Type I and Type II error.

Type I error is defined as an outcome that is classified as statistically significant when it is actually false in the population ("false positive"). Researchers have set a standard that a 5% false positive rate is acceptable. Most analyses report multiple statistical tests and as such, inflation of the Type I error rate should be considered. A Bonferroni correction adjusts the significance levels for the number of statistical tests by dividing the critical p value by the number of statistical tests performed [133–135].

		Significance	
		Significant	Non-Significant
Truth in the Population	True	Power $1-\beta$	Type II Error β
	False	Type I Error α	

Table 3.1: Summary of Type I and Type II Error

Type II error is defined as an outcome that is classified as statistically non-significant when it is actually true in the population ("false negative"). Statistical power is related to Type II error by $1-\beta$, which is the rate a statistically significant outcome is actually true in the population. If statistical power is low then there is a greater probability of making Type II error.

3.3 Analysis of Variance (ANOVA)

In the previous section, t-tests were used to compare means between two different groups; however, studies often have more than two groups. An analysis of variance, or ANOVA, is a statistical technique used to compare means across more than two groups of subjects. The statistic used to determine if the group means vary is called the F-statistic. The F-statistic takes the general form of

$$F = \frac{\text{explained variance}}{\text{unexplained variance}} \tag{3.5}$$

and is compared to critical values on the F-distribution.

The type of ANOVA depends on how the observations were collected. A between-subjects ANOVA is an analysis of variance where different subjects are observed at each level of the factors. A factor is a categorical variable identifying the groups of subjects measured. The sources of variation in a between-subjects ANOVA are the variability between the groups and the variability within the groups. A within-subjects, or repeated-measures ANOVA, is an analysis of variance where the same subjects are measured at each level of the factors. Because the same subject is measured in each group, a variability also exists between the means of each individual subject, a between-persons variability; this becomes an additional source of variability.

The four assumptions made prior to calculating an ANOVA are the same as the assumptions of a t-test [127]. However, a repeated-measures ANOVA has an important additional assumption: the subject's observations in each group are related. Both this assumption and the assumption of equal variances between groups is known as sphericity. A violation of sphericity can increase the probability of committing Type I error.

ANOVA can also be described by the complexity of the research design. For example, a one-way ANOVA compares means between groups of a single factor (e.g., High School Class Ranking). ANOVA can also be used to analyze means between groups of more than one factor (e.g., High School Class Ranking and Gender), a two-way ANOVA. Increasing the complexity of the research design increases the sources of variability. For example, in a two-way between subjects ANOVA, the between-group variability now stems from the variability of the first factor (main effect), the variability of the second factor (main effect), and the variability of the interaction between the first and second factor (interaction effect). Overall, ANOVA can be expanded to more than two factors (three-way ANOVA, etc.) but the type of ANOVA is always determined by the research design.

3.4 Linear Regression

Linear regression is a statistical method used to model the linear relationship between variables. Simple regression estimates a model of a continuous outcome with one single independent variable. However, the model can be extended to multiple regression, which estimates the model of a continuous outcome with more than one independent variable. Multiple regression considers the model of the form

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_{p-1} X_{i,p-1} + \epsilon_i \tag{3.6}$$

where Y_i is the value of the observed data for the ith participant, ϵ_i is the residual or error of the ith observation, β_j is the regression coefficient of the jth independent variable, and X_{ij} is the value of the ith observation of the jth independent variable [136]. The goal of multiple linear regression is to estimate β_j so that the error between the observed data and the estimated model is at a minimum; in other words, find the model which maximizes the variability explained in the continuous outcome by the multiple independent variables. The explained variability R^2 is characterized by

$$R^2 = 1 - \frac{\sum \epsilon_i^2}{\sum (Y_i - \bar{Y})^2}, \tag{3.7}$$

where $\bar{Y}$ is the mean of the observed data [136].

In regression, significance testing is also used to assess the strength of the relationship between two variables. Typically, the estimated regression coefficients are compared to a value of zero (no relationship between the dependent and independent variables). For example, if an estimated regression coefficient is statistically significant ($p < 0.05$) then the independent variable has a linear relationship with the outcome variable.

Regression analysis has five major assumptions: (1) linear relationship between variables, (2) all variables, including the residuals, are normally distributed, (3) no, or little, multicollinearity (an independent variable cannot be predicted by another independent variable), (4) the residuals are independent from each other, and (5) homoscedasticity (the variance of the residuals does not depend on the independent variables) [136].

One of the goals of linear regression is to estimate a model to maximize the explained variability in the outcome variable of interest. Hierarchical linear regression (HLR) is a regression technique to determine if the addition of an independent variable significantly improves the proportion of explained variance in the dependent variable. This technique adds one additional independent variable in a step-wise fashion and assesses the differences between the successive models.

3.5 Psychometric Theory

In general, psychometric theory, or test theory, is a collection of statistical models used to develop and evaluate the structure of an evaluation instrument [137]. The following sections will describe some methods used in test theory: Factor Analysis and Item Analysis, including Classical Test Theory, Item Response Theory, and Differential Item Functioning. Table 3.2 summarizes the commonly used statistics within these paradigms.

Classical Test Theory (CTT)		
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 $D < 0.2$
α	Cronbach's alpha	Values in $[0, 1]$; $\alpha > 0.7$ indicates acceptable reliability [138].
ϕ	Phi coefficient	Pearson correlation effect: 0.1 small, 0.3 medium, 0.5 large
Item Response Theory (IRT)		
b	Item difficulty	Typical range of -4 (easiest) to 4 (hardest)
a	Item discrimination	Typical range of -4 (least discriminating) to 4 (most discriminating)
V	Cramer's V	Goodness-of-fit; 0.1 small misfit, 0.3 medium, 0.5 large
Differential Item Functioning (DIF)		
$\Delta\alpha_{MH}$	Mantel-Haenszel	$\lvert\Delta\alpha_{MH}\rvert < 1$, negligible; $[1, 1.5)$, small to moderate; > 1.5, large
L	Lord's statistic	$\lvert L\rvert < 1$, negligible; $[1, 1.5)$, small to moderate; > 1.5, large
Confirmatory Factor Analysis (CFA)/Structural Equation Modeling (SEM)		
Model-Fit Indicies		Kline's Rule of Thumb [139]: $\chi^2/df \leq 3$
		Comparative Fit Index: CFI ≥ 0.95
		Root-Mean-Square Error of Approximation: RMSEA ≤ 0.05
		Standardized Root-Mean-Square Residual: SRMR ≤ 0.08

Table 3.2: Summary of commonly used statistics in Psychometric Theory.

3.5.1 Validity and Reliability

Determining the validity (accuracy) and reliability (precision) of an instrument is essential prior to discussing the various statistical procedures used in the development of an instrument. Validity is defined as "empirical evidence [...] to support the use of test scores for a stated purpose" [140]; in other words, whether the instrument measures what it is intended to measure. There are three types of validity discussed in instrument development: construct validity, content validity, and criterion validity. Construct validity refers to the degree to which the empirical evidence supports the theoretical purpose of the instrument. When researchers are discussing a "valid" instrument, they are typically referring to construct validity. Content validity refers to an examination of the overall content of an instrument; the extent to which the instrument covers all of the necessary content related to the intended construct that is to be measured. Criterion validity refers to the level of correlation between the measured construct of the instrument and a measure that is valid and should be related to that construct. The most important type of validity is construct validity because, if not identified, content and criterion validity will not exist.

Reliability is defined as “an indication of the consistency, stability, or precision of scores” [141]. Cronbach’s α is a common measure of reliability and will be discussed in section 3.5.2. It is important to note that although scores may be reliable, the instrument may not be valid; however, a test cannot be valid unless the scores are reliable. The following sections will discuss statistics used to assess the validity and reliability of an instrument.

3.5.2 Classical Test Theory

Classical Test Theory (CTT) is an important component of test theory. Under the assumption that test score T is equal to the sum of a hypothetical true score X plus random error E, the goal of CTT is to evaluate item and test reliability and validity [140]. Through an analysis of item difficulty, item discrimination, and correlations, CTT provides a basic way to evaluate the structure, reliability and validity of a test.

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 higher the item difficulty, the easier the item) [140]. Item discrimination D measures how well an item can distinguish between students who have strong knowledge of the subject matter from those who do not. Discrimination is defined as

$$D = P_u - P_l, \tag{3.8}$$

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 [140].

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." The validity framework for evaluating concept inventories proposed by Jorion *et al.* suggests items with $D < 0.2$, $P < 0.2$, or $P > 0.8$ as problematic for distractor-driven instruments [142–144].

In addition to item difficulty and item discrimination, overall test reliability and inter-item correlations are typically analyzed in CTT. Under the assumption of unidimensionality, Cronbach's alpha (α) is calculated to provide a measure of internal consistency reliability of an instrument [138, 141]. It is calculated using

$$\alpha = \frac{k}{k-1}\Big(1 - \frac{\sum_1^k \sigma_k^2}{\sigma_{Total}^2}\Big), \tag{3.9}$$

where k is the number of items, $\sum \sigma_k^2$ is the sum of the item variances, and σ_{Total}^2 is the variance of the total scores [141]. Cronbach's α of 0.7 is considered acceptable reliability and 0.9 as excellent [138]. Cronbach's α is sensitive to test length; as test length increases, α should increase as well. An item is problematic and should be eliminated if α increases with the removal of that item [142].

The relationship between the responses of two different items is called the inter-item correlation [140]; a Pearson correlation coefficient ϕ is commonly used. This statistic is calculated as

$$\phi_{jk} = \frac{p_{jk} - p_j p_k}{\sqrt{p_j q_j p_k q_k}}, \tag{3.10}$$

where p_j is the proportion of individuals answering item j correctly, $q_j = 1 - p_j$ is the proportion of individuals answering item j incorrectly, and p_{jk} is the joint proportion of individuals answering both item j and k correctly [140]. If the correlation between two different items is negative, elimination of this item should be considered.

To determine if two independent groups of individuals on the same dichotomously scored item are statistically significantly different, a χ^2 test of independence on a two-by-two contingency table of correct and incorrect answers for each population can be analyzed. A measure of effect size can then be calculated by $\phi = \sqrt{\frac{\chi^2}{N}}$, where N is the total number of individuals. The criteria for this effect size are $\phi = 0.1$ is a small effect, $\phi = 0.3$ is a medium effect, and $\phi = 0.5$ is a large effect.

3.5.3 Item Response Theory

Item Response Theory (IRT) is a system of models used to predict a binary outcome with a continuous latent variable, typically a characteristic of an individual such as ability [145, 146]. A latent variable is a variable that is not directly observed. Because the *test* is the unit of analysis, CTT ignores the repeated-measures nature of an instrument. IRT, however, treats the individual *item* as a unit of analysis to allow for item and person parameters to be measured on the same metric. Therefore, an individual's ability can be interpreted relative to the item and not just relative to the other individuals in the sample [137].

The goal of IRT is to model the probability of success for person i on item j, in terms of the individuals' latent ability trait and item parameters. Because of the binary outcome variable, IRT models take the form of a logistic function [145, 146]. There are three commonly used IRT models: the one-parameter logistic model (1PL) and the related Rasch model, the two-parameter logistic model (2PL), and the three-parameter logistic model (3PL). Many other models exist.

In a 1PL model, the probability of success π_{ij} is modeled in terms of the individuals' ability θ_i, the item difficulty b_j, and a constant item discrimination a. The 1PL model is

$$\pi_{ij} = \frac{\exp[a(\theta_i - b_j)]}{1 + \exp[a(\theta_i - b_j)]}. \tag{3.11}$$

Item difficulty is defined as the value of θ_i at which an individual has a 50% chance of success [145, 146]. Item discrimination is the slope of the tangent line at the item's difficulty; the degree to which an item discriminates between an individual with knowledge and one without knowledge [145, 146]. If the discrimination is constrained to one, $a = 1$, the 1PL is called the Rasch model [147]. The model that is most closely related to CTT is the 2PL model. This model assumes that each item has a discrimination a_j and a difficulty b_j:

$$\pi_{ij} = \frac{\exp[a_j(\theta_i - b_j)]}{1 + \exp[a_j(\theta_i - b_j)]}. \tag{3.12}$$

The 3PL model adds an additional parameter, called the guessing parameter c_j. This parameter is defined as the value of π_{ij} when θ_i approaches $-\infty$. In other words, if an individual has extremely low ability, the probability of success is solely reliant on guessing. The 3PL model is

$$\pi_{ij} = c_j + (1 - c_j)\frac{\exp[a_j(\theta_i - b_j)]}{1 + \exp[a_j(\theta_i - b_j)]}. \tag{3.13}$$

IRT has two important assumptions: (1) unidimensionality and (2) local independence [145, 146]. Unidimensionality assumes that a single latent trait is being measured; this assumption provides evidence for construct validity. Local independence assumes that the responses to an item are independent of any other item. These assumptions tend to come hand in hand because if only one trait is being measured then an individual's response only depends on the single trait and nothing else.

IRT introduces a model for student responses and therefore model fit should be assessed. A common way to investigate model fit is to divide the individuals into G groups and then compare the predicted mean of the group, given by the IRT model, with the observed mean of the group [148–150]. An effect size is typically reported along with goodness-of-fit. The model fit is evaluated with a χ^2 distribution with $df = G - n$, where n is the number of parameters in the IRT model. An effect size, Cramer's V, can also be calculated as $V = \sqrt{\frac{\chi^2}{(df \cdot N)}}$. The item characteristic curves (the plots of the logistic functions) should still be examined for all items.

In addition to model fit, the certainty of the estimated parameter can be calculated for IRT models; the certainty is known as information [145]. In general, since the error in an item's estimated parameter can be calculated, the amount of information is defined as

$$I = \frac{1}{\sigma^2}, \tag{3.14}$$

where σ^2 is the variance of the item parameters. In other words, the more precise the item parameter is (i.e., the smaller the variability in the estimated parameter), the more information is given by the value of the parameter. Typically in IRT, the item information $I_i(\theta)$ function is computed for each individual item i and the test information is the sum of all item information functions. The test information function provides the amount of information given from the overall test at any level of ability θ.

3.5.4 Differential Item Functioning

In test development, the concept of instrumental bias toward a certain group of individuals is typically explored. More recently, the focus of instrumental bias has revolved around item-level statistics rather than overall differences between group means. Differential Item Functioning (DIF) analysis is a technique to detect items that are functioning differently across different groups of individuals [145]. 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. For this manuscript, DIF statistics will be used to compare difficulty parameters across two groups: male and female students. The DIF statistics will only focus on differences in difficulty parameters rather than both the difficulty and discrimination parameters.

A common measure of DIF is the Mantel-Haenszel (MH) statistic [151–153] which has been employed by the Educational Testing Service (ETS) for 25 years to examine item fairness in high-stakes exams [154]. 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 female students to the odds of answering an item i correctly for male students:

$$\alpha^i_{MH} = \frac{\sum_{k=0}^{S} \frac{N^{CM}_{ik} N^{IF}_{ik}}{N_k}}{\sum_{k=0}^{S} \frac{N^{CF}_{ik} N^{IM}_{ik}}{N_k}}, \tag{3.15}$$

where N^{CM}_{ik} is the number of male students with total score k who answered item i correctly, N^{IM}_{ik} is the number of male students with total score k who answered item i incorrectly, N^{CF}_{ik}

is the number of female students with total score k who answered item i correctly, N_{ik}^{IF} is the number of female students with total score k who answered item i incorrectly, S is the total score on the instrument, and N_k is the total number of students with score k [155]. While relatively complicated, if only one group was formed this reduces to the odds ratio of men answering the item correctly to women answering correctly,

$$\frac{odds^M}{odds^F} = \frac{p^M(1-p^F)}{p^F(1-p^M)} \tag{3.16}$$

where p^M is the probability male students answer the item correctly and p^F is the probability female students answer the item correctly. Eqn. 3.15 generalizes Eqn. 3.16 by separating the odds ratios by total test score.

DIF statistics can also be calculated in terms of the item parameters measured by IRT. Many statistics have been constructed; however, this manuscript will report Lord's statistic which compares the difference in difficulty parameters between male and female students with the average difference in difficulty [156]. This statistic uses the difficulties estimated from the Rasch model. The Lord's statistic L_i is

$$L_i = b_i^{\mathrm{F}} - b_i^{\mathrm{M}} - \frac{1}{n}\sum_{j=1}^{n}(b_j^{\mathrm{F}} - b_j^{\mathrm{M}}), \tag{3.17}$$

where b_i^{F} is the item difficulty of the female students on item i, b_i^{M} is the difficulty of the male students on item i, and n is the number of items on the instrument.

The two DIF statistics presented can be transformed into an effect size $\Delta\alpha_{MH}$. α_{MH} is transformed to $\Delta\alpha_{MH}$ by $\Delta\alpha_{MH} = -2.35\ln(\alpha_{MH})$ [154]. Lord's statistic can be transformed by multiplying L_i by 2.35 [151, 155]. Male students have an advantage when the DIF statistic is less than zero and female students have an advantage when the DIF statistic is greater than zero. The ETS classifies a DIF statistic with magnitude less than 1 as negligible DIF, a DIF statistic with magnitude between 1 and 1.5 as small to moderate DIF, and a DIF statistic with magnitude greater than 1.5 as large DIF [157].

3.5.5 Factor Analysis

Factor analysis is a method used to assess the internal structure of an instrument [158, 141]. In instrument development, two types of factor analysis are examined: exploratory factor analysis (EFA) [159–161] and confirmatory factor analysis (CFA) [161, 162]. Both of these analytic techniques are important to establishing the validity of an instrument.

Exploratory Factor Analysis

Following the initial data collection, EFA is used to analyze the dimensionality of the instrument [141]. This method assumes no prior relationship between the items and the measured constructs (factors), hence, the name exploratory. This implies that all variables are allowed to correlate with every factor. The goal of EFA is to maximize the explained variance shared among the individual items with the smallest number of factors [163, 141].

The most common method to extract the appropriate number of factors in EFA is called principal component analysis (PCA) [164], which tries to explain the variance in the set of items. A factor solution is estimated by removing the maximum amount of variance explained from the original correlation matrix between items and in turn, leaving a residual correlation matrix. This process is continued with the new residual correlation matrix until all of the variability in the set of items is explained [141]. In PCA, the eigenvalues (proportion of variance accounted for by each factor) of the correlation matrix are analyzed and a decision is made on how many factors to extract. A common method to assess how many factors to extract is called parallel analysis [141]. This method compares the eigenvalues from the sample data to the eigenvalues estimated from a random dataset [160]. A large number of factor selection methods exist.

The other decision researchers have to make when performing an EFA is the type of rotation to use. Factor rotation refers to the alignment of the reference axes of the extracted factors [165, 141]. There are many types of rotation: varimax rotation and promax rotation are used in this manuscript. Varimax rotation assumes the extracted factors are orthogonal, or uncorrelated, and promax rotation assumes that the extracted factors are correlated.

Confirmatory Factor Analysis

In instrument development, EFA and CFA are typically used hand-in-hand. After EFA, CFA is used to confirm an a priori model proposed by the researcher and, therefore, gives the researcher more control [141]. The a priori model involves the hypothetical constructs that the instrument was trying to measure. In a standard CFA model, each item can be explained by a single factor and measurement error, which is independent of any other item's measurement error. Factor loadings, which are simply regression coefficients between a latent variable and an observed variable, are then estimated and the model-fit is assessed.

A χ^2-statistic and Kline's rule of thumb for good model fit $\chi^2/df \leq 3$ [139] are typically reported. However, the weaknesses of a 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 [139] have led to a number of additional statistics with superior performance and extensive research into combinations of statistics [166]. This continues to be an active area of research and general rules for model fit are still under development [167].

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 [166] 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.

Structural Equation Modeling

Structural Equation Modeling (SEM) is an extension to CFA where a combination of multiple linear regression models and confirmatory factor analysis models are analyzed [168]. This allows for complex data to be modeled with multiple observed and latent variables.

SEM begins with model specification, followed by model identification, model estimation, model testing, and model modification. Model specification involves all of the relevant theoretical variables of interest. Model identification, estimation, and testing uses the same techniques as confirmatory factor analysis; χ^2, CFI, RMSEA, and SRMR are examined. Model modification is the technique used to remove parameters that are not significant in the fitted model. This process is done by iteratively removing the parameters with the worst statistical significance followed by refitting and comparing the two nested models. The removal of a parameter must also make sense in the theoretical framework of the study. SEM allows for more advanced models to be analyzed when trying to explain a theoretical framework.

The research presented in this manuscript used these statistical methods to answer the relevant research questions.

Chapter 4

Learning Assistants in the Introductory Laboratory Setting[*]

This chapter will present a case study of the adoption of a Learning Assistant model. The results of this analysis were the motivation to explore the gender gap within the conceptual inventories in the future chapters of this manuscript.

[*] This work was supported by the National Science Foundation under Grant No. EPS-1003907.

4.1 Introduction

Substantial evidence suggests that replacing traditional instruction with methods that engage students in their own learning can produce substantial improvements in the conceptual mastery of physics [30]. However, the adoption of reform teaching methods is incomplete and many instructors abandon these methods after trying them [169]. The purpose of this chapter is to explore one possible implementation of reformed teaching strategies that might insulate student conceptual learning from the changes in course personnel brought on by the growth of the number temporary teaching faculty. This section will also examine the implementing department's experiences with the reform methods to shed light on the decision to abandon a reform.

4.2 Research Questions

- R1: Can an implementation of a Learning Assistant (LA) program in the introductory laboratory setting using the *Tutorials in Introductory Physics* effectively support conceptual learning and improved attitudes toward science?
- R2: Is its effectiveness independent of the instruction in the lecture part of the course? Is the effectiveness dependent on the individual teaching assistants and learning assistants?
- R3: Is it equally effective for all student populations?
- R4: What is the minimum dataset required to understand the effectiveness of the program?

Beyond answering these questions, this chapter will present a case study of a well-resourced attempt to adapt some of the most widely deployed PER pedagogies to the local educational environment and the struggles of the department to understand and manage the outcomes of the program. The program was ultimately abandoned because, while it showed positive outcomes, these outcomes were clinically relatively weak and not identified in time to secure funding for the program's continuation.

4.3 Implementation

The research was conducted at a large eastern land-grant university serving approximately 30,000 students. The LA program, modeled after the program developed at the University of Colorado [36], was implemented in both the introductory, calculus-based mechanics class (Physics 1) and the introductory electricity and magnetism class (Physics 2). This research will focus on the educational impact of the program on the students in the reformed classes.

The impetus for the reform was a department culture that valued teaching and a growing departmental awareness of the proven effectiveness of reform teaching methods. This impetus was catalyzed into action by the funding of a large five-year traditional science grant that had a substantial educational component supporting the reform. Although non-lecture-based structures have been proven to be more effective in physics instruction [170, 171], there was insufficient space and financial resources available to support a studio format. To maintain consistent lab space and TA staffing levels, an implementation of the University of Washington *Tutorials in Introductory Physics* [31] presented by LAs in the laboratory environment was enacted.

Both Physics 1 and 2 were presented with four hours of lecture per week (four days per week) in large lecture theaters. Students also co-enrolled in a two-hour, once-per-week laboratory class. This course structure was not changed through the implementation of the LA program. The lecture was presented by 14 different instructors over the project; these instructors established their own lecture pedagogy, pacing of lecture, homework and examination policies. The lecture portion of the course was often not well timed to support the laboratory; with labs investigating material not yet covered in lecture. For most instructors, their teaching of the course was decoupled from the student's laboratory experience. Lecture instructors were of varying standings within the department from late-career graduate students to tenured professors and had a wide range of teaching experience.

Prior to the introduction of the LA program, students performed traditional two-hour cookbook experiments in the laboratory supervised by graduate teaching assistants (TAs). Students completed lab reports on the experiments in a graded lab notebook and completed weekly laboratory quizzes. The lab reports and lab quizzes were graded by the TAs. Students received a lab grade that formed 10% of their course grade.

After the introduction of the LA program, the first hour of the laboratory was dedicated to small group work using the *Tutorials*; the LA acted as lead lab instructor during this time. The LA received training in the general pedagogy of engaged science instruction during a one-credit class taught by an expert in science education during his or her first LA semester. This science pedagogy course was also modeled after the University of Colorado Learning Assistant course [36]. The elements of the Colorado model were recreated to the extent possible. Specific instruction on the presentation of the upcoming week's tutorial was provided by the program lead in a weekly meeting separate from the pedagogy class. During this meeting, the LAs worked the upcoming *Tutorial*, were encouraged to reflect on their previous week's teaching experience, and received pedagogical advice about the instruction of next week's *Tutorial.* In the second hour of the lab, the students worked in the same lab group on a traditional laboratory experiment with the TA acting as the lead lab instructor. The homework associated with the *Tutorial* was assigned as lab homework; the laboratory exercise in the second hour also had associated homework questions. Students' written laboratory and tutorial homework were graded by the TA. Students continued to receive a grade for the laboratory that accounted for 10% of their course grade. The LA was assigned to three laboratories per week and was required to hold office hours. All laboratory sessions received LA support. Each LA was compensated with a stipend ($1,500/semester). The university has a flat tuition model, and therefore the additional pedagogy course credit did not cause the students to incur additional costs. The LAs were much more gender balanced than the student or TA populations in the classes (Students: 80% male, TAs: 87% male, LAs: 64% male).

4.4 Methodology

This research was conducted between the spring 2011 and spring 2015 semesters in the introductory, calculus-based physics sequence. In Physics 1, the LA program was introduced in the fall 2011 semester using the spring 2011 semester as a control semester. In Physics 2, the program began in the fall 2012 semester using three control semesters, spring 2011, fall 2011, and spring 2012.

Conceptual learning was evaluated with the FMCE [22] in Physics 1 and with the CSEM [23] in Physics 2. Both examinations were taken in lecture as a pretest and a post-test. Students received credit for good faith efforts on each examination. Attitudes towards science were measured using the Colorado Learning Attitudes about Science Survey (CLASS) [79], also taken as a pretest and a post-test within the lecture environment. This instrument was designed to measure how well students' self-reported beliefs and attitudes about physics aligned with "expert" beliefs and attitudes about physics. The CLASS was scored as suggested by the developers by measuring the number of questions on which students expressed an attitude similar to those of experts (favorable responses) or responses opposite to those of experts (unfavorable responses). The change in both the number of favorable and unfavorable responses between the pretest and post-test will be reported.

Only students who completed the course, completed both the conceptual pretest and post-test, completed the CLASS pretest and post-test, and who had reported a SAT or ACT score to the university were included in the study. During the period studied, 3,531 students completed Physics 1. Of these students, 3,180 had reported ACT or SAT scores, of these, 2,263 completed both the FMCE pretest and post-test, and of these, 2,025 also completed the CLASS pretest and post-test. These $N = 2,025$ students form the Physics 1 sample. For Physics 2, 2,507 students completed the class, of these, 2,230 reported ACT or SAT scores, of these, 1,357 completed the CSEM pretest and post-test, and of these, 1,454 completed the CLASS pretest and post-test. These $N = 1,454$ students form the Physics 2 sample.

Over the nine semesters studied, class enrollment for Physics 1 averaged 390 students per semester and Physics 2 averaged 280 students per semester. Both classes consisted of 80% male students, 20% female students, 85% engineering majors and 15% non-engineering

Instructor	Standing	Involvement	PER	Physics 1 (N)	Physics 2 (N)	Physics 1 Semesters	Physics 2 Semesters
1	TF	Y	Y	1168	199	1, 3, 4, 5, 6, 7, 8	9
2	F	N	N	38	0	1H	
3	F	N	N	41	0	1	
4	A	N	N	95	0	2	
5	F	N	N	50	320	2	1, 3, 5
6	F	N	N	44	0	3H, 5H	
7	A	N	N	86	66	4	2
8	A	N	N	410	387	6, 8, 9	7, 8
9	A	N	N	35	0	7H	
10	A	Y	Y	206	40	7, 9, 9H	8H
11	F	Y	Y	0	103		2H, 4H, 6H
12	A	Y	N	0	214		2, 4
13	F	Y	N	0	147		6, 7
14	A	Y	N	0	121		6

Table 4.1: Instructor standing codes: F=Tenure/Tenure-track Faculty, TF=Permanent teaching faculty, A=Temporarly instructors. The code H indicates honors sections.

majors. Honors sections of the classes were offered during the spring semesters for Physics 1 and during the fall semesters for Physics 2. Honors students had substantially higher standardized test scores and received instruction in smaller classes. Because they represent a distinct, and relatively small population, they were excluded from the analysis for this study.

A total of 14 different instructors of various standing from full professors to late career graduate students taught lecture sections of the courses during the period studied. The program lead was also the lecture instructor for Physics 1 or 2 during some semesters (Instructor 1 in the analysis). These classes were offered with two to three lecture sections each semester. All instructors were invited to participate in the LA training sessions, but only some accepted. Many lecture instructors taught multiple lecture sections in a semester. The teaching assignments and some properties of the instructors are summarized in Table 4.1. Instructors were classified by their standing as either tenure/tenure-track faculty, permanent teaching faculty, and temporary instructors which included adjuncts, late career graduate students, and research professors. Instructors who attended the LA training sessions are marked as “Involved.” Instructors with significant interest or knowledge of PER as denoted as “PER.” Only the program lead (Instructor 1) taught in both pre-LA and post-LA semesters in Physics 1. Instructors 5, 11, and 12 taught in pre- and post-LA sections in Physics 2, but Instructor 11 taught exclusively honors students.

The efficacy of this implementation of the *Tutorials in Introductory Physics* using LAs in the laboratory setting was investigated using the normalized gain on the FMCE in Physics 1 and the CSEM in Physics 2. The normalized gain was calculated and converted to a percentile to aid in interpretation of regression coefficients. Students’ demographic characteristics were characterized using a number of variables: Gender (Female=0, Male=1), Race/Ethicity (Non-Caucasian/Hispanic=0, Caucasian Non-Hispanic=1), and whether the student was a first generation college student (FirstGen=0 for non-first generation students, FirstGen=1 for first generation students). The sample was not sufficiently diverse to explore race and ethnicity more fully. The student’s general ability was characterized using their score on the ACT or Scholastic Aptitude Test (SAT). Both the SAT Mathematics and Verbal scores were used. For the ACT, the Mathematics subscore and English subscore were used. In both cases, these scores were converted to percentile ranks using the guidelines published by the testing companies. For students reporting both ACT and SAT scores, the average of the two percentile scores was used. The resulting variables will be designated ACT/SAT Math and ACT/SAT Verbal. The analysis will identify a strong difference between spring and fall semesters showing spring semesters with stronger results in Physics 1 and fall semesters with stronger results in Physics 2. This spring/fall variation may have resulted from students taking Calculus 1 in their entering freshman semester matriculating to Physics 1 in their spring, freshman semester and then to Physics 2 in their fall, sophomore semester.

These students were calculus-ready upon entering the university and were matriculating following the prescribed plan of their departments; they were "on sequence." To capture this effect in the analysis, the variable "Fall Semester" was introduced (Fall Semester=1 for fall semesters, Fall Semester=0 for spring semesters). Students are on-sequence in Physics 1 if Fall Semester= 0 and on-sequence in Physics 2 if Fall Semester= 1.

4.5 Results

Table 4.2 shows the overall averages for Physics 1 and Physics 2. The difference between the average of the male students and the female students is presented in parentheses (a positive number means that the male students had a higher average than the female students). A t-test was performed for the gender difference in each quantity; the significance of the differences are presented as superscripts. A t-test aggregating all non-honors students in each class showed that there was no statistically significant difference between male and female students in their incoming ACT/SAT Math percentile scores or in the final course grade for either class. In both classes, female students had a significantly higher ACT/SAT Verbal percentile.

Overall, Table 4.2 shows a small increase in post-test scores in Physics 1 with the introduction of the LA program accompanied by a somewhat larger, but still small increase in normalized gain. Physics 2 also showed a small increase in post-test scores and normalized gains. CLASS scores degraded with the introduction of the LA program in both classes.

	Physics 1		Physics 2	
N (Male, Female)	Control Semesters 162 (127, 35)	LA Semesters 1863 (1509, 354)	Control Semesters 321 (257, 64)	LA Semesters 1133 (944, 189)
Pretest (%)	$29 \pm 19(5.5)$	$27 \pm 17(5.8)^c$	$25 \pm 10(2.8)^a$	$26 \pm 10(4.2)^c$
Post-test (%)	$47 \pm 25(1.0)$	$53 \pm 26(11)^c$	$35 \pm 16(5.1)^b$	$42 \pm 16(7.9)^c$
Normalized Gain (%)	$26 \pm 32(-0.7)$	$39 \pm 34(11)^c$	$13 \pm 17(4.1)$	$20 \pm 18(6.0)^c$
ACT/SAT Verbal	$71 \pm 18(-13)^c$	$69 \pm 20(-5.7)^c$	$72 \pm 20(-9.0)^b$	$69 \pm 20(-3.7)^a$
ACT/SAT Math	$81 \pm 15(-3.4)$	$77 \pm 16(.10)$	$80 \pm 15(-2.7)$	$78 \pm 16(2.0)$
Course Grade	$3.1 \pm .94(-.38)$	$3.0 \pm .94(.05)$	$3.1 \pm 0.91(-.17)$	$3.1 \pm 0.83(-.09)$
CLASS Fav. Change	$-2.4 \pm 17(-4.9)^a$	$-2.8 \pm 17(2.0)^a$	$-2.8 \pm 16(.78)$	$-4.5 \pm 16(.14)$
CLASS Unfav. Change	$2.4 \pm 13(4.2)$	$3.9 \pm 14(-2.3)^b$	$3.3 \pm 14(.40)$	$3.7 \pm 13(-.06)$

Table 4.2: Physics 1 and Physics 2 Averages. The number in parentheses represents the difference between the average for male students and female students. The superscript indicates whether the gender difference is significant. Superscript "a" denotes $p < 0.05$, "b" $p < 0.01$, and "c" $p < 0.001$.

4.5.1 Preliminary Analysis

Initially, ANOVA was used to explore the effect of the program and any differential effects on students of different gender, race/ethnicity, or first generation status. In Physics 1, the LA program was a significant treatment effect for the normalized gain on the FMCE $[F(1, 2023) = 20.57, p < 0.001]$. A two-way ANOVA with the LA treatment and gender showed both to be significant main effects ($p < 0.001$). The interaction between LA and gender was not significant. An interaction plot is shown in Figure 4.1. While the LA program did have a positive effect on male students, it had very little effect on female students. A two-way ANOVA using race/ethnicity and the LA program found significant main effects of the LA program ($p < 0.001$) and race/ethnicity ($p < 0.001$) but no significant interactions. An ANOVA for first generation status and the LA program found the main effect was significant for the LA treatment

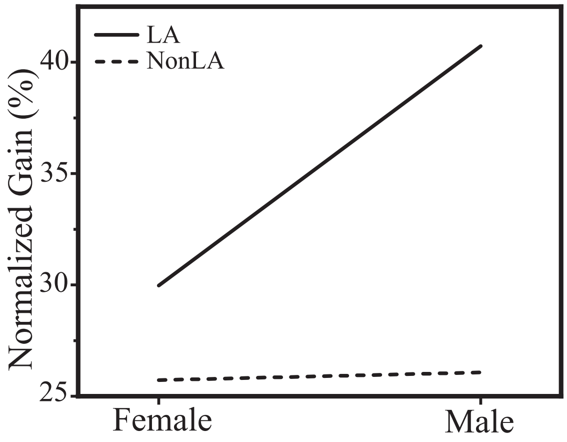

Figure 4.1: Interaction Plot of LA Program and Gender for Physics 1 FMCE post-test percentage.

$[F(1, 2021) = 20.58, p < 0.001]$; however, the first generation status main effect and the interaction were not significant.

The results for Physics 2 were somewhat different. The LA program was still a significant treatment effect on the normalized gain on the CSEM $[F(1, 1452) = 45.73, p < 0.001]$. Like Physics 1, both the LA program and gender were significant main effects, but their interaction was not significant. The main effect of race/ethnicity was not significant but the LA program main effect was significant $[F(1, 1450) = 45.85, p < 0.001]$; the interaction was also significant $[F(1, 1450) = 4.35, p = 0.04]$. The LA program was a significant main effect $[F(1, 1450) = 45.93, p < 0.001]$ in a two-way ANOVA with first generation status, but the first generation status was not. The interaction between the LA program and first generation status was a significant interaction $[F(1, 1450) = 5.27, p = 0.02]$. Exploration of this interaction showed the normalized gain of first generation students decreased with the introduction of the LA program.

Many of these findings were counter to what was expected by the program and could not be explained by the replacement of cookbook laboratories with well-vetted PER developed materials and providing well-trained, near-peer, more gender balanced support of these activities. Further, the apparent differential effect of the program on male and female students was unacceptable; the department could not support a curricular reform that preferentially benefited the overrepresented population.

4.5.2 The Role of Ability

To shed light on the initial observations, alternate explanations were explored. The normalized gain scores on the FMCE in Physics 1 exhibited a strong variation between

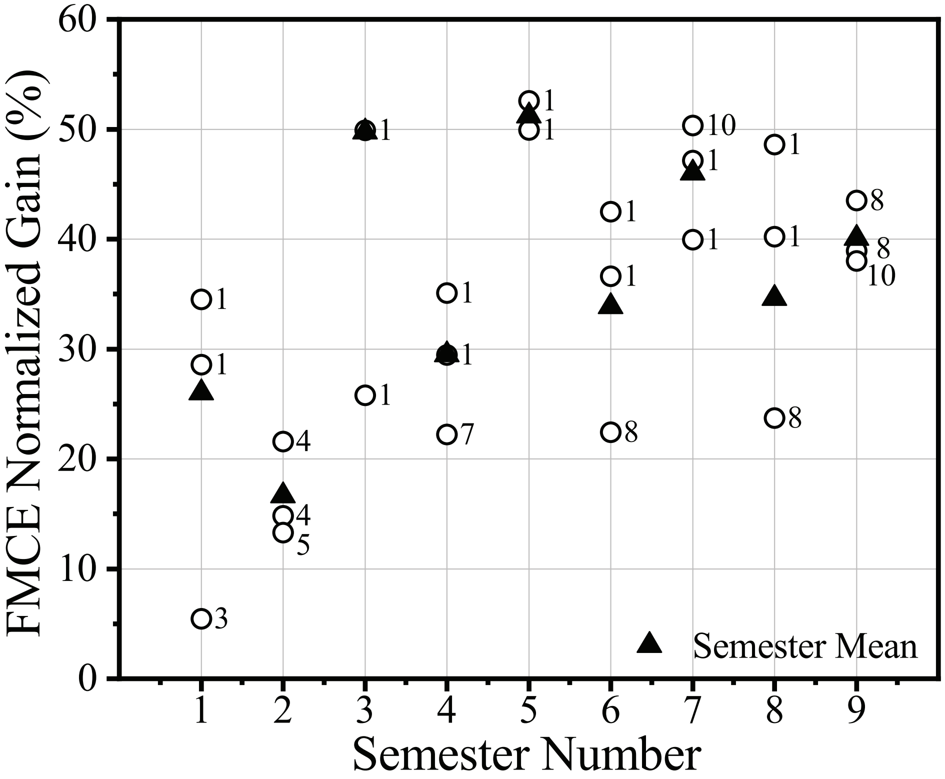

Figure 4.2: FMCE Normalized Gain vs. Semester in Physics 1. The open circles are independent lecture sections numbered by instructor. The triangle represents the semester mean.

spring (odd numbers) and fall (even numbers) semesters as shown in Figure 4.2. While Table 4.2 shows a relatively small increase in normalized gain comparing all LA semesters to the control semester, the increase in normalized gain considering only the spring (odd) semesters was more pronounced. The average normalized gain for the spring LA semesters was 47% as compared to 27% for the control semester. The LA program was presented in the same format and managed by the same program lead in each semester, and therefore this variation in conceptual learning outcomes could not be explained in terms of the instruction the LAs presented.

Focusing on only the spring semesters did not explain the gender disparity identified in the previous section. In the control semester, the normalized gain of male students was 28% which rose to 50% after this implementation of the LA program. For female students,

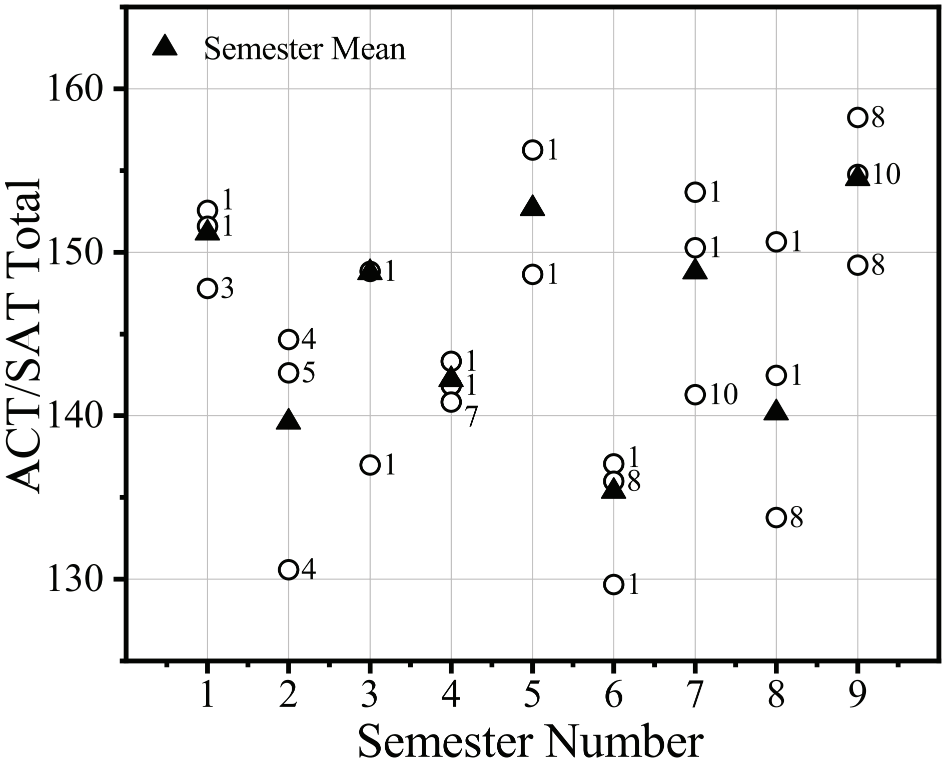

Figure 4.3: ACT/SAT Total vs. Semester in Physics 1. The open circles are independent lecture sections numbered by instructor. The triangle represents the semester mean.

the control semester produced normalized gains of 27% which rose to only 35% with the LA program.

The standardized testing scores of the students also varied substantially from spring to fall. ACT/SAT Math and ACT/SAT Verbal percentiles were summed to form the total standardized test score, ACT/SAT Total; this variable had a maximum walue of 200. ACT/SAT Total is plotted by semester in Figure 4.3. The figure shows the same pattern of spring/fall semester variation as the normalized gain scores. Figure 4.4 plots the normalized gain vs. the ACT/SAT Total for semesters after the implementation of the LA program and shows a correlation between the variables. The central line represents the regression line for all instructors. Instructor 1 taught many of the lecture sections in the sample and on average out-performed other instructors; Instructor 1 is represented by the upper line and

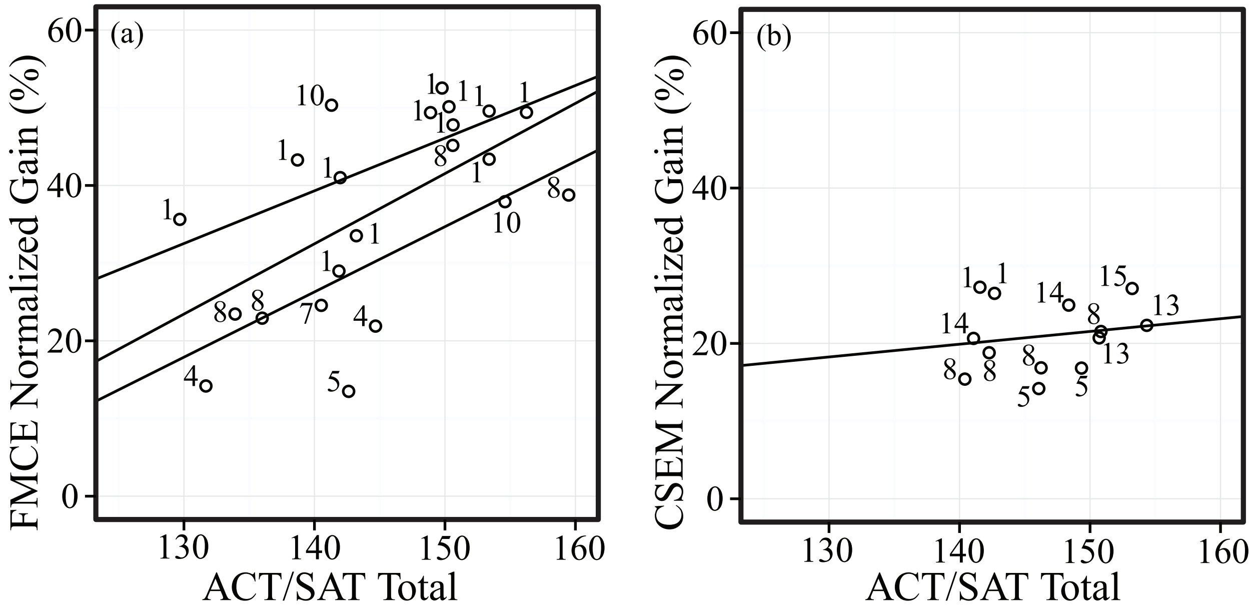

Figure 4.4: Normalized Gain vs. ACT/SAT Total in (a) Physics 1 and (b) Physics 2 for semesters after the implementation of the LA program. In (a), the top line represents Instructor 1, the center line all instructors, and the lower line instructors other than Instructor 1. In (b), the line represents all instructors.

all other instructors the lower line. While Instructor 1 was, in general, the most successful, other instructors were also successful with some outperforming the ability-corrected gains of Instructor 1.

Physics 2 exhibited the same pattern of oscillation of ability and normalized gain scores as did Physics 1, except that while the spring semesters were the high ability semesters in Physics 1, the fall semesters were the high ability semesters in Physics 2. This adds support to the interpretation of the oscillation as an on-sequence/off-sequence effect. The Physics 2 data demonstrated a weaker trend when normalized gain was plotted against ACT/SAT Total as shown in Figure 4.4. The weaker linear relationship was influenced by the both the more restricted range of CSEM results and the range of the ACT/SAT Total percentile. Multiple regression results will find both ACT/SAT percentile scores were significant variables for both Physics 1 and 2, with a somewhat weaker effect in Physics 2.

A relationship between ACT/SAT Total percentile and FCI [21] normalized gain had previously been reported by Coletta, Phillips, and Steinert [172]. The correlation r of ACT/SAT Total with normalized gain was calculated. Aggregating all semesters (including the control semesters) yielded FMCE, $r = 0.37$ [$t(2023) = 17.97, p < 0.001$], CSEM, $r = 0.32$ [$t(1452) = 12.75, p < 0.001$]. If only the LA treatment semesters are included these values become FMCE, $r = 0.39$ [$t(1861) = 18.07, p < 0.001$], CSEM, $r = 0.33$ [$t(1131) = 11.88, p < 0.001$]. The FMCE correlations are similar to the $r = 0.46$ reported by Coletta *et al.* for university students using the FCI; the CSEM correlations are somewhat smaller.

4.5.3 Regression Analysis

Hierarchical linear regression analysis allowed the exploration of the many variables associated with the program while controlling for the variability in student ability and demographic composition. Variables were added to regression models as shown in Table 4.3. Regression coefficients are reported: unstandardized (B) and standardized (β). Model fit was characterized with the adjusted-R^2 (R^2_{adj}) and SE is the standard error of B.

At each step, the new model was tested using ANOVA to determine if the added variables explained significantly larger variance. Only models demonstrating significant improvement are reported. Most reported steps produced a model which ANOVA identified as explaining additional variance at the $p < 0.001$ level of significance. The effect of the fall semester was weaker in Physics 2 and the addition of this variable produced a model that significantly improved on the previous model but with a significance level of $p = 0.02$.

After Model 2, models were fit that added the race/ethnicity or the first generation

		Physics 1 (n=2,025)				Physics 2 (n=1,454)			
		B	SE	β	R^2_{adj}	B	SE	β	R^2_{adj}
Model 1	ACT/SAT Math	0.56^c	0.06	0.26^c	0.14^c	0.24^c	0.04	0.21^c	0.11^c
	ACT/SAT Verbal	0.26^c	0.04	0.15^c		0.13^c	0.03	0.14^c	
Model 2	ACT/SAT Math	0.53^c	0.06	0.25^c	0.16^c	0.22^c	0.04	0.20^c	0.12^c
	ACT/SAT Verbal	0.31^c	0.04	0.18^c		0.15^c	0.03	0.16^c	
	Gender	11.95^c	1.80	0.35^c		6.38^c	1.19	0.35^c	
Model 3	ACT/SAT Math	0.47^c	0.06	0.22^c	0.17^c	0.21^c	0.04	0.19^c	0.12^a
	ACT/SAT Verbal	0.31^c	0.04	0.18^c		0.15^c	0.03	0.17^c	
	Gender	12.01^c	1.78	0.35^c		6.31^c	1.19	0.35^c	
	Fall Semester	-8.44^c	1.44	-0.25^c		2.17^a	0.90	0.12^a	
Model 4	ACT/SAT Math	0.47^c	0.06	0.22^c	0.20^c	0.22^c	0.03	0.19^c	0.16^c
	ACT/SAT Verbal	0.31^c	0.04	0.18^c		0.16^c	0.03	0.18^c	
	Gender	11.77^c	1.76	0.34^c		6.04^c	1.17	0.33^c	
	Fall Semester	-11.16^c	1.46	-0.33^c		1.68	0.89	0.09	
	LA Treatment	19.60^c	2.60	0.57^c		8.27^c	1.06	0.45^c	
Model 5	ACT/SAT Math	0.45^c	0.05	0.21^c	0.24^c	0.21^c	0.03	0.19^c	0.19^c
	ACT/SAT Verbal	0.32^c	0.04	0.19^c		0.17^c	0.03	0.19^c	
	Gender	12.17^c	1.72	0.35^c		6.26^c	1.14	0.34^c	
	Fall Semester	-6.43^c	1.62	-0.19^c		3.05^a	1.44	0.17^a	
	LA Semester	15.90^c	2.95	0.46^c		5.38^c	1.48	0.30^c	
	Instructor	$(-26, 0)^c$	(5.42, 0)	$(-0.77, 0)^c$		$(-11, 0)^c$	(1.72, 0)	$(-0.62, 0)^c$	
Model 6	ACT/SAT Math	0.45^c	0.05	0.21^c		0.21^c	0.03	0.19^c	
	ACT/SAT Verbal	0.31^c	0.04	0.19^c		0.17^c	0.03	0.19^c	
	Gender	12.16^c	1.72	0.35^c		6.24^c	1.14	0.34^c	
	Fall Semester	-6.08^c	1.60	-0.20^c		2.81^a	1.34	0.15^a	
	LA Semester	16.80^c	2.87	0.49^c		5.75^c	1.41	0.32^c	
	Instructor SD	*9.49*		*0.28*		*4.00*		*0.22*	
Model 7	ACT/SAT Math	0.35^c	0.05	0.17^c	0.26^c	0.21^c	0.03	0.19^c	0.19^c
	ACT/SAT Verbal	0.28^c	0.04	0.17^c		0.17^c	0.03	0.19^c	
	Gender	9.78^c	1.70	0.28^c		6.42^c	1.16	0.35^c	
	Fall Semester	-6.08^c	1.59	-0.18^c		2.98^a	1.44	0.16^a	
	LA Semester	16.91^c	2.90	0.49^c		5.47^c	1.48	0.30^c	
	Pretest Percent	0.37^c	0.04	0.18^c		-0.04	0.04	-0.02	
	Instructor	$(-24, 0)^c$	(5.33, 0)	$(-0.69, 0)^c$		$(-11, 0)^c$	(1.72, 0)	$(-0.61, 0)^c$	
Model 8 Restricted Sample	ACT/SAT Math	0.28^c	0.05	0.13^c	0.23	0.14^c	0.04	0.13^c	0.22
	ACT/SAT Verbal	0.14^c	0.04	0.08^c		0.11^c	0.03	0.12^c	
	Gender	6.47^c	1.50	0.19^c		4.08^c	1.23	0.222^c	
	Fall Semester	-8.39^c	1.50	-0.24^c		7.19^c	1.63	0.39^c	
	LA Semester	11.83^c	2.98	0.34^c		5.36^b	1.84	0.29^b	
	Pretest Percent	0.14	0.15	0.07		-0.23	0.13	-0.13	
	Instructor	$(-25, 0)^c$	(4.90, 0)	$(-0.73, 0)^c$		$(-13, 0)^c$	(2.64, 0)	$(-0.74, 0)^c$	

Table 4.3: Hierarchical Linear Regression Model Predicting Normalized Gain. Gender was coded with male as one, female as zero. Normalized gain was expressed as a percentage. In Model 5, 7, and 8, the range of instructor coefficients are presented with Instructor 1 set to zero. Superscript “a” denotes $p < 0.05$, “b” $p < 0.01$, and “c” $p < 0.001$.

status of the students. These additional variables did not produce a significantly improved model. Additional models were explored building on Model 6 by the addition of the LA or TA as a random effect; these additions did not significantly improve model fit. Figure 4.2 shows Instructor 3 in Semester 1 to be a potential outlier; removing this instructor had very little effect of the regression coefficients.

After the optimal model, Model 5, was identified, interaction terms were added for the LA treatment, gender, race/ethnicity, and first generation status. No combination of these interaction terms produced a model that was statistically significantly superior to Model 5. The apparent interaction between the LA treatment and gender was lifted with the addition of the other variables.

Model 5 was fit with the instructor as a fixed effect and Model 6 with the instructor as a random effect. The results for Models 5, 7, and 8 in Table 4.3 report the range of coefficients for the individual instructors measured against the effect of Instructor 1, which was set to 0. The parameter estimate in Model 6 for the instructor is the standard deviation of the distribution of intercepts. Bootstrapping showed that Model 6 was also a significant improvement over Model 4 ($p < 0.001$). There are features of the data that suggest both models. The relatively large number of course sections suggest the instructor can be treated as a random variable; however, the wide range of outcomes suggest the data does not fully explore instructor differences. The fixed and random effects models produced very similar regression coefficients.

Regression models were also produced for the post-test percentage with similar results with the LA treatment producing a significant regression coefficient of 8.53 for the LA Semesters for Model 5 in Physics 1 and 6.19 for Physics 2. This represents the amount the

LA treatment increased post-test scores controlling for the other variables in Model 5.

Overall, focusing on either Model 5 or 6, the LA program produced an increase of normalized gain of 16% in Physics 1 and a smaller increase of 5% in Physics 2. The difference was smaller if the coefficients were normalized producing a 0.5 standard deviation increase in Physics 1 and a 0.3 standard deviation increase in Physics 2. The inclusion of the ACT/SAT variables was significant, but did not fully explain the spring/fall variation. There was a substantial, significant effect of gender in all models; this effect was not explained by differential abilities of the male and female populations. In fact, as shown in Table 4.2, female students actually outperformed male students on their overall course grade. Neither race/ethnicity nor first generation status was significant in predicting normalized gain after controlling for ACT/SAT scores. While the LA program improved student conceptual learning, it was not sufficient to insulate students from the effect of the lecture instructor. The weakest performing lecture instructors had the effect of lowering student learning gains more than the LA program could raise them.

4.5.4 The Role of Gender

The regression results in Table 4.3 show gender as a consistently significant regressor. This result has been reported elsewhere and has been extensively explored as discussed in the introduction. One explanation considered has been that pre-preparation, measured by the pretest score, could explain the difference in conceptual performance. Pretest score is used to calculate normalized gain and contains a measure of facility with conceptual physics problems and, as such, is of concern as an explanatory variable. Including the pretest percentage in Model 7 did lower the effect of gender somewhat in Physics 1 but did not eliminate it; it had little effect in Physics 2. Model 7 was a statistically significant improvement on Model 5 ($p < 0.001$) for both classes. Because pretest scores also show gender differences, it is unclear whether this result is informative about the differences in performance.

To examine whether pre-preparation in physics could explain the different outcomes of students of different gender, a sub-sample of students who scored less than 25% on the pretest was examined in Model 8. Female students compose 23% of this subgroup while composing only 13% of the sub-group with pretest scores greater than 25%. The sub-sample contained $1,189$ students. Model 8 shows focusing only on students who demonstrated little preparation in the subject decreased the effect of gender, but did not eliminate it. Pretest score was not a significant regressor for this model.

A similar analysis was carried out for Physics 2 using a pretest threshold of 25%; this produced a sub-sample containing 665 students of which 23% were female; female students composed only 15% of the students scoring over 25%. Again, the effect of gender was reduced but not eliminated. Female students' under-representation in the group of students with higher pre-preparation accounts for some but not all of the differential performance on both the FMCE and CSEM.

4.5.5 CLASS Results

Hierarchical linear regression was repeated for the CLASS. Models 4 and 5 in Table 4.3 were analyzed for both the change in Favorable attitude toward scientist-like thinking and the change in Unfavorable attitude toward scientist-like thinking. In Physics 1, the regression model explained little of the variance in these variables. Model 4, with the change in CLASS Favorable responses as the dependent variable, produced $R^2 = 0.00$ [$F(5, 2000) =$

$1.01, p = 0.40$] and the change in Unfavorable attitudes produced $R^2 = 0.00$ [$F(5, 2000) = 2.95, p = 0.02$]. While the model for change in Unfavorable attitude was significant, only the ACT/SAT variables were significant regressors. The LA treatment was not a significant regression variable for either dependent variable. Adding Instructor as a fixed effect, Model 5 produced models that explained substantially more variance: change in Favorable attitude $R^2 = 0.02$ [$F(11, 1994) = 5.04, p < 0.001$] and change in Unfavorable attitude $R^2 = 0.02$ [$F(11, 1994) = 4.84, p < 0.001$] with many instructors as significant regression coefficients. The unstandardized beta coefficients for the different instructors ranged from -11.18 to 0 for the change in Favorable and 0 to 7.58 for the change in Unfavorable with Instructor 1 as the 0 in each case. As such, the effect of the LA program on students' scientist-like thinking was very small, much smaller than the differential effects of the range of lecture instructors. ANOVA confirmed that Instructor was a significant treatment effect for both variables in Physics 1: change in Favorable attitudes [$F(6, 1994) = 8.38, p < 0.001$] and change in Unfavorable attitudes [$F(6, 1994) = 6.59, p < 0.001$].

For Physics 2, Model 4 was not statistically significant for either Favorable or Unfavorable changes (Favorable $R^2 = 0.00$, Unfavorable $R^2 = 0.00$). Model 5 including the instructor as a fixed effect was significant, but more weakly so than in Physics 1: change in Favorable attitudes $R^2 = 0.01$ [$F(11, 1440) = 2.90, p < 0.001$] and change in Unfavorable attitudes $R^2 = 0.01$ [$F(11, 1440) = 2.42, p < 0.01$]. Instructor was the only significant effect in Physics 2 for both dependent variables. The range of unstandardized beta coefficients for the instructors, with Instructor 1 as the zero reference, was from -6.87 to 2.4 for Favorable change and -1.21 to 4.81 for Unfavorable change. ANOVA confirmed that Instructor was a significant treatment effect in Physics 2 as well: Favorable [$F(6, 1438) = 4.44, p < 0.001$]

and Unfavorable $[F(6, 1438) = 3.92, p < 0.001]$.

4.5.6 Instructor Effects

Table 4.1 shows that the instructors of the classes varied in their level of involvement in the LA program, their standing within the department, and their PER knowledge. Different instructors also had greatly different conceptual learning gains and CLASS changes; the effect of the lecture instructor was often much larger than the effect of the LA program. Only three instructors were classified as having strong PER knowledge, the program lead, Instructor 1 (permanent teaching faculty), Instructor 10 (temporary faculty), and Instructor 11 (tenure-track faculty). Instructor 11 taught only honors students and was not included in this analysis; however, he or she achieved comparably excellent results with a normalized gain of 33% in Physics 2. Instructors 1 and 10 produced excellent results with normalized gains of 43% and 44% in Physics 1, respectively. Permanent teaching faculty (Instructor 1) and temporary faculty outperformed tenure-track faculty in Physics 1 with temporary faculty achieving gains of 29% while tenure-track faculty achieved gains of 9%. If Instructor 10 is removed from the analysis, temporary faculty achieved gains of 24%. The low gains by tenure track faculty in Physics 1 may be influenced by small sample size; only 91 students included in the sample were taught by tenure-track faculty. In Physics 2, Instructor 1 produced a normalized gain of 25%, tenure-track faculty a gain of 17% and temporary faculty 20%. For Physics 2 tenure-track faculty that actively involved themselves in the LA program by coming to the LA training session achieved normalized gains of 23% while actively involved temporary faculty achieved gains of 22%. The number of instructors, the distribution of PER knowledge, and involvement in the program make general conclusions difficult, but it does not appear that instruction by non-tenure-track faculty degraded the learning experience in these classes.

4.6 Discussion

This study sought to investigate four research questions which will be discussed in the order proposed.

R1: Can an implementation of an LA program in the introductory laboratory setting using the Tutorials in Introductory Physics effectively support conceptual learning and improved attitudes toward science? The implementation of an LA program using the *Tutorials in Introductory Physics* was partially successful and showed some promise as a mechanism to expose students to research-based materials in an instructional environment where a large number of instructors rotate through the lecture component of the course and where some, but not complete, support of reformed education exists. The implementation generated statistically significant improvements in normalized gain in both Physics 1 and 2 with an increase of 16% in Physics 1 and 5% in Physics 2 (Model 5); this represents 0.5 standard deviations in normalized gain in Physics 1 and 0.33 standard deviations in Physics 2. While a significant effect, the FMCE post-test score of 53% and the CSEM post-test score of 42% suggest that there is substantial additional room for improvement.

R2: Is its effectiveness independent of the instruction in the lecture part of the course? Is the effectiveness dependent on the individual teaching assistants and learning assistants? The program was unsuccessful at eliminating the effect of the lecture portion of the course on conceptual learning. The effect of the lecture instructor was commensurate or larger than the effect of the LA program in many cases. Neither the individual teaching assistant, nor the learning assistant, was a significant random effect in the regression models.

R3: Is it equally effective for all student populations? While there appeared to be an interaction between the gender of the student and the LA program in Physics 1, this interaction ceased to be significant when the ability of the student measured by ACT/SAT percentile and whether the student was on-sequence were added to the models. No effect of race/ethnicity or first generation status was identified. The LA program improved the conceptual learning of all populations equally; however, it was not sufficient to close a significant gap in conceptual learning between male and female students. The gender interaction identified in the preliminary analysis was the result of the female students in the control semester of Physics 1 having higher than average ACT/SAT scores which translated into higher normalized gain scores on the FMCE. Once ability was controlled, the interaction disappeared.

R4: What was the minimum dataset required to understand the effectiveness of the program? The variables used in Model 5, ACT/SAT Math and Verbal percentile, gender, and spring/fall semester were the minimum set of descriptive control variables needed to understand the program. Preliminary analysis without these variables produced spurious results that were later contradicted when more complete models were investigated. To accumulate the spring/fall control variable, a minimum of two consecutive semesters of control data should have been collected. Although race/ethnicity and first generation status were not significant variables in this study after controlling for ability, it would be remiss to exclude them from future analysis.

The failure to achieve the same level of improvement in Physics 2 as was observed in

Physics 1 requires further examination. In both classes, the LAs were prepared through the same process which was overseen by the project leader. Both classes performed the same number of *Tutorial* exercises which were graded using the same policy; as such, both LA treatments were extremely similar. We hypothesize, but can offer little additional support beyond the similarity of the treatments, that the actual learning outcomes were equivalent. The second semester course covers many topics that are poorly represented on the CSEM: Gauss' Law, Ampere's Law, circuits, and optics. All of these important topics were however covered by the *Tutorials*. The weak performance of students on the CSEM may have resulted from a mis-match of the topics covered in the *Tutorials* and the conceptual coverage of the CSEM.

The analysis used to investigate the program identified a significant relation between normalized gain and student ability measured by the ACT/SAT Math and Verbal percentiles. The relation of FCI normalized gain scores and ability had been reported [172]; this study serves to confirm that similar effects exist in other conceptual instruments. The strong correlation of ability and normalized gain suggests that, when comparing the same educational intervention at different institutions or in different semesters at the same institution, student ability should be used as a control variable.

Controlling for ability failed to fully explain the large differences between the spring and fall semesters. The spring/fall variation involves factors not fully accounted for by standardized test scores. As such, the Fall Semester variable should be included in analyses comparing multiple semesters.

This study provided additional evidence of a differential performance by gender on conceptual instruments even when male and female students performed equivalently on other course measures such as test scores. The effect of gender on normalized gain was of the same size as the LA treatment.

The desired increases in student attitudes measured by the CLASS and the positive effects on underrepresented populations weren't realized. In fact, CLASS scores degraded, but not significantly.

4.7 Case Study

The purpose of this study was not only to describe the results of implementing two of the most well-vetted PER innovations in a laboratory setting, but also to describe the department's experience using PER materials. This experience ultimately resulted in the department replacing the LA/Tutorial program with another PER inspired set of materials and a restricted implementation of LAs in the laboratory setting where the LA earned course credit instead of receiving monetary compensation. The results did not allow the department to argue for the continued funding of the LA program, nor did they argue for the continued use of the *Tutorials* which added to the already heavy cost of textbooks for the students.

The LA program was part of a larger five-year funded scientific project. This encouraged the department to persevere when initial results were discouraging. The LA program was initially implemented in Physics 1, and therefore the first feedback on the program received was the normalized gains shown in Figure 4.2. The net effect of the LA program in Semester 2 appeared to be a substantial reduction in normalized gain.

Further, until Semester 9, the last semester of the project, it appeared that only Instructor 1 and 10 could produce exemplary results; Instructor 1 was the program lead and

Instructor 10 taught only in the high ability spring semesters. Instructor 10 had shadowed Instructor 1's lecture presentation and presented it with little modification. It was only when Instructor 8 moved from lower performing fall semesters to a higher performing spring semester in Semester 9 that it became clear that other instructors less closely related to Instructor 1 could achieve significant gains for the higher achieving population. Instructor 8 reported no change in his or her lecture presentation in Semester 9.

The results in Physics 2 were clinically mediocre. While the regression results ultimately demonstrated an increased normalized gain, the post-test scores did not suggest a clinical effect that could be interpreted as substantially improving student learning. With the data from Physics 1 being initially negative and then mixed, and Physics 2 producing limited gains it is unlikely the department would have sustained the program past its initial year were it not part of a larger funded project.

The program only became fully understood with the analysis of Semester 9; previous to this semester the department hypothesized that the primary effect on conceptual learning was the exemplary lecture presentation of Instructors 1 and 10. By this time, substantial plans for the modification of the program post funding (beginning Semester 10) were in place and no work had been initiated to secure additional funding for the LA program.

4.8 Implications for Instruction

The implementation of a Learning Assistant program using the *Tutorials in Introductory Physics* in the laboratory setting was successful for students in introductory, calculus-based mechanics. It was less successful for students in introductory, calculus-based electricity and magnetism. Implementing the program in the laboratory setting was not sufficient to inoculate students from other factors in the course design outside of the laboratory; substantially different outcomes were identified for different lecture instructors. As such, it is critical to have coherent support across all course elements for lab-based reforms.

4.9 Implications for the Adoption of Research-based Materials

Overall, the department's experience with LAs presenting the *Tutorials in Introductory Physics* curriculum was mixed. Their persistence was largely a result of the program's role as part of a larger funded project. Because of the fluctuation in student ability, the initial results of the project (Semester 2 in Physics 1) indicated that it was having a negative effect on learning. The project also seemed to help female students less than male students. In general, a few guidelines for the adoption of a research-based curriculum would have made the program more successful and allowed the department to solicit funds to continue the program:

1. The research-based materials were introduced into a traditional course structure that was not working well. The project would have proceeded very differently if the courses had first been modified to produce coordinated instruction between different instructors in the same semester, the instructors and the laboratory, and instructors over multiple semesters. This would have allowed each instructor to support the value of conceptual learning with their examinations and made sure each student had the information needed to get the most out of lab.

2. The program used some of the most well-known products of PER, but it is unclear if the

curricular elements (the *Tutorials*) were well aligned with the assessment elements (the FMCE and the CSEM). The department would have benefitted from a packaging of elements so that all worked well together. We note that there are assessments aligned with the *Tutorials*, but these are not multiple-choice instruments and could not be deployed for classes of the sizes at the implementing institution.

3. The department underestimated the number of control semesters and the amount of background data necessary to understand the program. The variables in Model 5 represent the minimum background/demographic information and a minimum of two semesters of control data should have been collected.

4. The assessment instruments (the FMCE and CSEM) showed the program needed some modification, but did not give any indication of what that modification should be. Assessments designed to help educators fine-tune educational reform could have transformed the project.

4.10 Implications for Research Methodology

The normalized gain was introduced to allow comparison of educational data across different student populations. The results of this study support previous work that showed it is only partially effective for this purpose and that additional data is required to compare student outcomes for populations with substantially different academic preparation and ability. This study also strongly suggests the need for multiple semesters of control data. This study also shows that ability data in itself is not enough to fully account for the differences in student populations between the spring and fall semesters. The observation that normalized gain is correlated with student ability measures may suggest a reevaluation of the normalized gain results at institutions that feature an academically diverse population or studies that compare the normalized gain between different institutions.

4.11 Limitations

This work was performed at a single institution and, therefore, only the effect of the range of ability and preparation available at that institution was explored. While 14 instructors were included in the dataset, this cannot be viewed as a fully representative sample of all instructors, and therefore, the range of the effect of instructor is probably understated.

4.12 Future Work

The differences in the normalized gain score by gender without corresponding differences in course grade have been reported previously [124]; however, this work extended those findings to show there was also no underlying differences in ability measured by ACT/SAT score. The differences by gender were also not explained by the student matriculating on the preferred sequence for graduation, taking the class in the correct semester. While student pre-preparation was shown to explain some of these gender differences additional work is needed to fully understand this effect and how the differences can be moderated.

4.13 Conclusion

A Learning Assistant program using the *Tutorials in Introductory Physics* in a traditional laboratory setting showed significant conceptual learning gains in introductory calculus-based physics classes at a large land-grant university. These learning gains were significantly related to student ability (ACT/SAT score), gender, and whether the student was "on-sequence." Controlling for ACT/SAT scores, no significant relation with either race/ethnicity or first generation status was identified. While the program was implemented in the laboratory setting, the learning gains were significantly affected by the lecture instructor.

Chapter 5

Exploring the Gender Gap in the Conceptual Survey of Electricity and Magnetism[*]

[*] This chapter presents the work published in Physical Review Physics Education Research [61]. This work was constructed with collaborative efforts from Adrienne Traxler, Lynnette Michaluk, and Gay Stewart. It 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 TOPP: Taxonomy of Physics Problems grant, DUE-0535928.

5.1 Introduction

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

5.2 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?

5.3 Methods

5.3.1 Context for Research

The research was conducted in the second-semester, calculus-based physics course at a large southern 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 [173]. At the time of the analysis, the undergraduate engineering program was ranked 105th [174]; 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 Chapter 2 was performed at more highly ranked institutions. For example, the University of Colorado-Boulder’s undergraduate engineering program was ranked 32 and Colorado School of Mines 44 at the time the data was accessed [174]. 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 took the course because it was required for their major.

While there was some spring/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 5.1.

Women were substantially underrepresented in the course for all semesters studied.

Students were required to attend two 50-minute lectures each week and two two-hour 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 post instruction; 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 [175].

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

Table 5.1: Class size and gender composition by semester.

The course studied was both designed to be 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.

5.3.2 Identifying Non-Quantitative Problems

Each problem presented in the course was classified as either quantitative or non-quantitative 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 non-quantitative. The identification of non-quantitative 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 3μ 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 3μ 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.

5.3.3 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 (non-quantitative 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 student’s 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 post-instruction 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.

5.3.4 Sample

The data was collected from the fall 2007 semester to the spring 2012 semester. During this time, 1,851 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 $N = 1,407$ students which 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 dataset 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.

5.3.5 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. This work will employ 15 statistical tests. A Bonferroni correction would adjust significance levels with $p < .05$ becoming $p < .0033$, $p < .01$ becoming $p < .00067$, and $p < .001$ becoming $p < .000067$. Few results will be changed by this correction. Most tests produced significance levels of $p < .001$; these results were also significant at the $p < .000067$ 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 SEM analysis and the many statistical tests implied by the analysis were treated as independent and not included in this correction.

5.4 Results

Table 5.2 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 qualitative problems on the in-semester tests (3%). Male and female students performed equally on in-semester quantitative test problems. The gender differences were examined using t-tests. Significant differences between male and female students were found

Problem Collection	Male Students $(M \pm SD)\%$	Female Students $(M \pm SD)\%$
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

Table 5.2: Male and female student averages for different problem collections.

on the CSEM pretest [$t(729) = 8.59$, $p < .001$, $d = .46$], the CSEM post-test [$t(531) = 5.92$, $p < .001$, $d = .37$], qualitative laboratory quiz problems [$t(508) = 3.37$, $p < .001$, $d = .22$] and qualitative test problems [$t(495) = 2.80$, $p = .005$, $d = .19$]. 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 $p < .05$ level if corrected for the number of statistical tests performed using a Bonferroni correction.

The dataset was reduced from the 1,851 students who completed the course for a grade to the 1,407 student sample for this study by the restriction to students who completed all problems on both the pretest and post-test. If this restriction is relaxed, the pretest and post-test averages change little. For the 1,788 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 1,613 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 5.2. For the post-test, 1,665 students answered any question with an average percentage correct of 64.0% [men 65.2%; women 59.8%]. Of these, 1,582 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 5.2. Blank questions were treated as incorrect in this analysis.

5.4.1 The Effect of Pretest Score

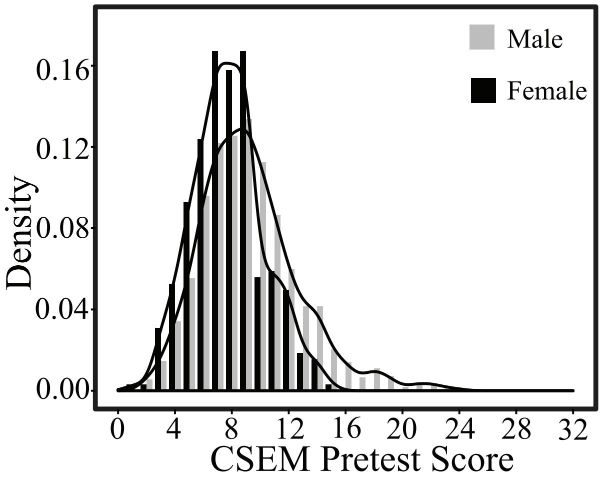

Figure 5.1: The distribution of CSEM pretest scores for male and female students.

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. 5.1. 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. Table 5.2 and Fig. 5.1 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. 5.2.

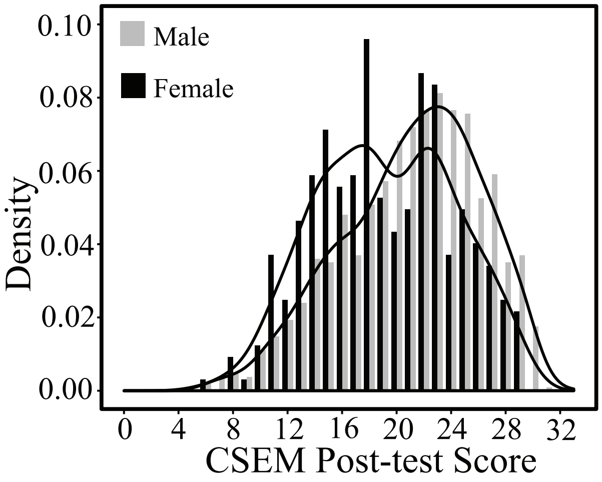

Figure 5.2: The distribution of CSEM post-test scores for male and female students.

To explore the effects of these differences in pretest scores on students' performance post instruction, 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 female students scored 13 or above on the pretest for analysis.

Figure 5.3 presents the average score within each pretest range for male and female students for each problem collection. 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. 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 CSEM post-test grew rapidly with 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. 5.3.

5.4.2 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 as plotted in Fig. 5.3. All have small differences at the lowest pretest scores, but a growing

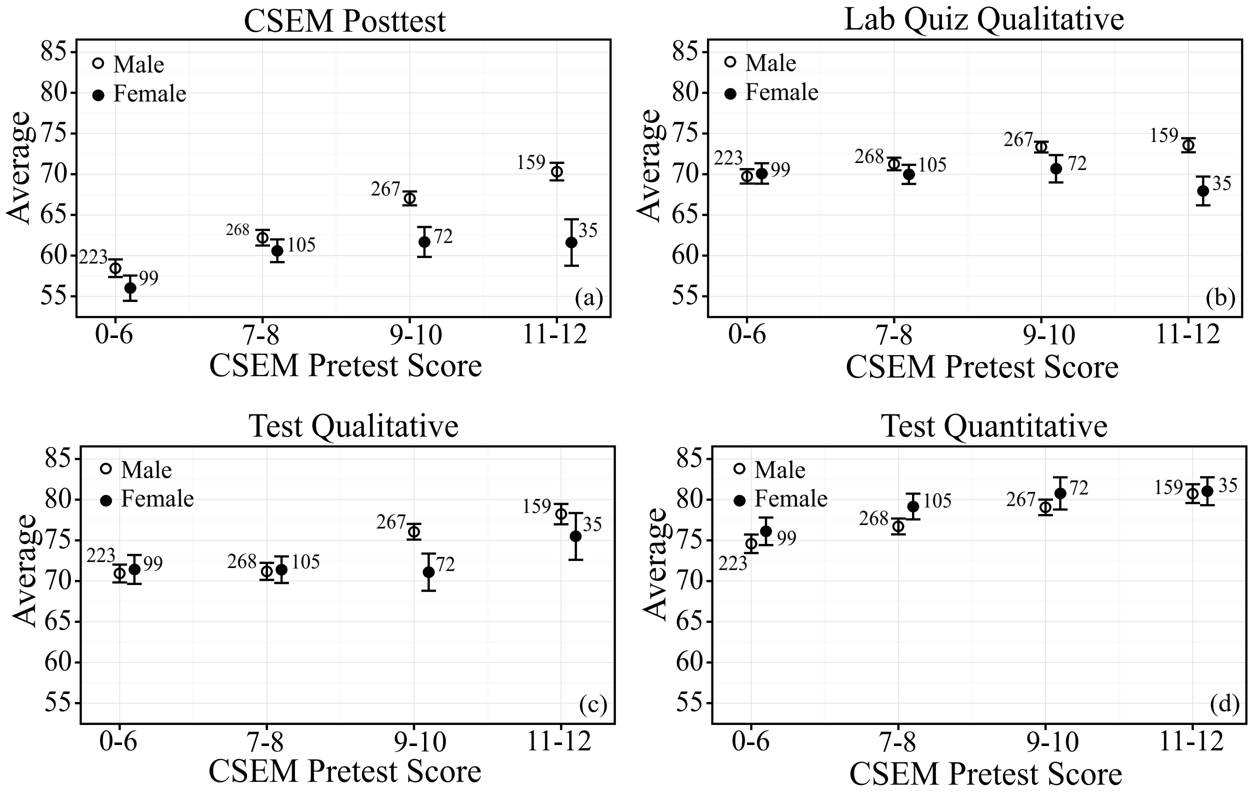

Figure 5.3: Evaluation Average vs. CSEM pretest: (a) the CSEM post-test, (b) qualitative lab quiz problems, (c) qualitative test problems, and (d) the quantitative test problems.

difference between male and female outcomes becomes apparent as 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 [176], 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 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. 5.4. 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.

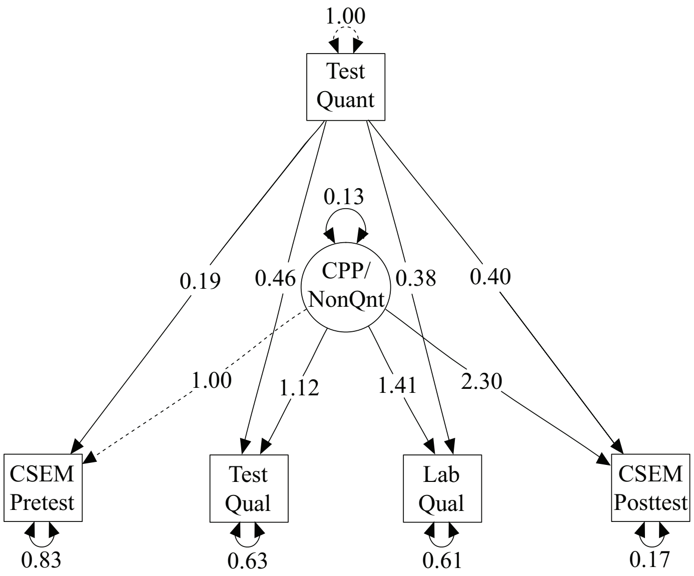

Figure 5.4: Structural Equation Model for CPP/NonQnt’s relation to qualitative problem performance.

The resulting model had generally good fit parameters. The chi-squared statistic, [$\chi^2(2) = 6.21$, $p = .045$], was on the border of that required for rejecting the null hypothesis of perfect model fit, and near Kline’s $\chi^2/df \leq 3.0$ [139] rule of thumb for good model fit.

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. All other criteria were well within the range of good model fit: RMSEA = .039, SRMR = .012, and CFI = .997. The 90% confidence interval of the RMSEA was .005 to .075. All regression coefficients, factor loadings, and variances were significant (*ps* < .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.

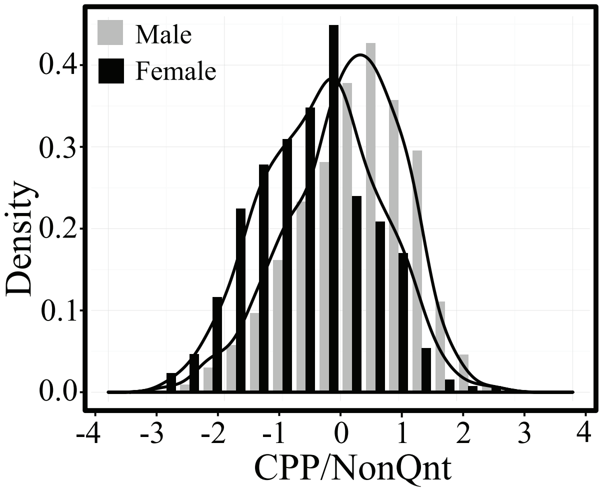

Figure 5.5: The distribution of male and female students' CPP/NonQnt.

The distribution of male and female CPP/NonQnt is shown in Fig. 5.5; 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. 5.5 was significant [$t(517) = 7.0$, $p < .001$; male students $M = .10$, $SD = 1.00$ and female students $M = -.34$, $SD = .98$]. 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, .44, represents a small to medium effect size.

The binning used in Sect. 5.4.1 was repeated in Fig. 5.6 which demonstrated a growing difference in CPP/NonQnt with CSEM pretest score, as well as an approximately linear relation between male pretest scores and CPP/NonQnt. 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 correlation r between between pretest score and CPP/NonQnt was smaller for women, $r = .20$ [$t(321) = 3.57, p < .001$], than for men, $r = .41$ [$t(1082) = 14.86, p < .001$]. 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.

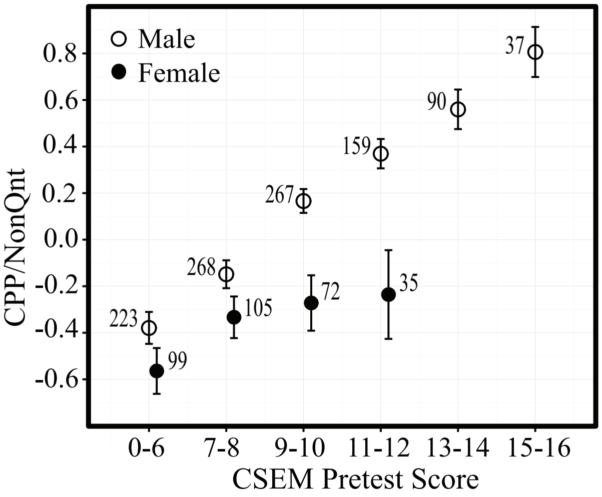

Figure 5.6: CPP/NonQnt vs. CSEM pretest score.

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 [$t(400) = 2.4, p = .018$]; 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 Fig. 5.3 and 5.6 are similar, their interpretation is quite different. Fig. 5.6, 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 5.7 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.

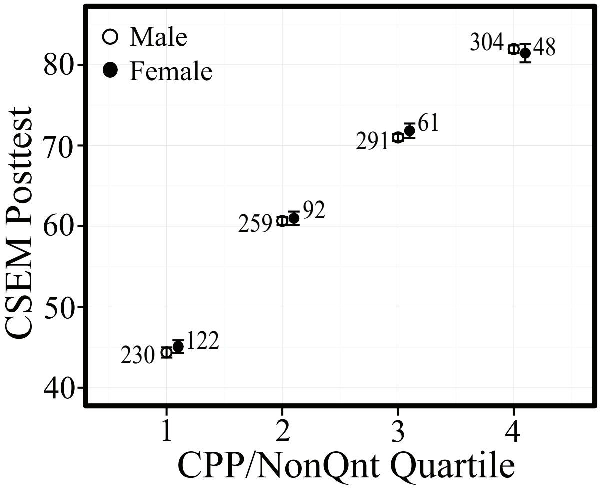

Figure 5.7: CPP/NonQnt Quartile vs. CSEM post-test score.

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 5.3.

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.

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

Table 5.3: Male and female student CPP/NonQnt by quartile.

5.4.3 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 5.1 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 Fig. 5.8(a) and (b), respectively.

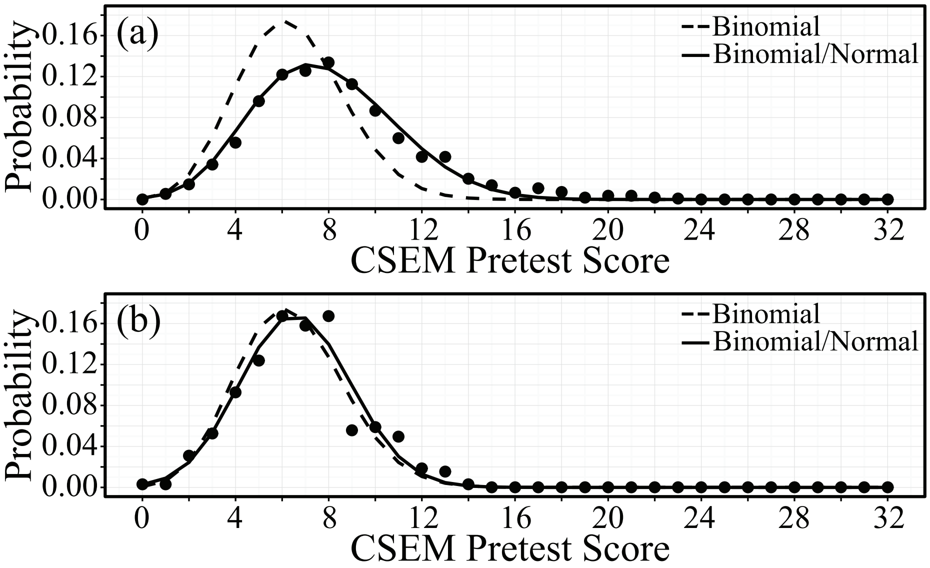

Figure 5.8: Model fits for the probability distributions of CSEM pretest scores for (a) male students and (b) female students.

The dashed lines in Fig. 5.8 show the result of fitting only a binomial distribution, $B(x; p = .2)$, representing pure guessing with probability of success $p = .2$ 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 dissimilar from the observed distribution for male students. The solid lines in Fig. 5.8 plot the result of fitting the model

$$P(x) = p_b \cdot B(x; p = .2) + p_n \cdot N(x; \mu_n, \sigma_n) \tag{5.1}$$

which 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.

Fitting Eqn. 5.1 with $p_b + p_n = 1$ yielded $p_b = .40$, $p_n = .60$, $\mu_n = 8.83$, and $\sigma_n = 2.99$ for the male students. For the female students the fit resulted in $p_b = .23$, $p_n = .77$, $\mu_n = 6.67$, and $\sigma_n = 2.36$. The curve representing Eqn. 5.1 substantially improves the fit to the male distribution of pretest scores, as shown in Fig. 5.8(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 the mean of the binomial guessing distribution, 6.40. The difference between the binomial and binomial/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. 5.6 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.

5.5 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 ($N = 2,000$) [74] of the BEMA, but inconsistent with the negative gender gap observed by Pollock ($N = 168$) [51]. The growth of the gender gap from pretest to post-test was consistent with Kohl and Kuo, but of a smaller magnitude [73]. 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 5.2 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 5.3 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 [74] where the gender gap disappeared if students were binned by FMCE post-test scores. It was also different than the CSEM normalized gain results of Kohl and Kuo who found a fairly consistent gender gap, except in the lowest pretest bin [73]. The growth of achievement gaps with increasing student ability has been well documented [86]; 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 the Kost-Smith, Pollock, and Finkelstein study should be substantially more academically prepared than those in this study; in fact, Kost-Smith, Pollock, and Finkelstein [74] 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 (Figure 5.8) 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? SEM 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 .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 [74, 75].

The CSEM pretest score was more weakly correlated with CPP/NonQnt for female students, $r = .20$, than for male students, $r = .41$. 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 the pure guessing distribution. If CPP/NonQnt rather than pretest score is employed to bin students no post-test gender gap exists (Fig. 5.7).

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 the introduction, 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 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 [93, 94], science anxiety [97], or stereotype threat [110] 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 [177–179] and the failure to detect a relationship between the fraction of women in a class and gender gaps [63]. 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 [117, 118], and in Chapter 6 we will show that CSEM items are generally fair. 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 [76]. 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 [87, 88] could impact some qualitative items, one would expect that female superiority at verbal reasoning [89] 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 [66] 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 [83]. Further research should be able to test this conjecture. This interpretation is not fully supported by the work of Kost-Smith, Pollock, and Finkelstein [74] 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 a 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 [83]. 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 non-interactive classes [30] as well as specific experiments investigating the effect of quantitative problem solving on conceptual learning suggest conceptual and quantitative learning are somewhat different processes [180].

5.6 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 accurate for women than for men [140, 142]; 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. 5.4.3.

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 in this dataset on average have a disadvantage in the 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 post instruction. 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 [181] 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 [182], 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 [65]. Self-efficacy has been demonstrated to be important in retention [183] and is one of the strongest psychological correlates with academic performance [184]; 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.

5.7 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.

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 [185]. Instructor gender effects were also observed in one of the course sections in the Kost-Smith *et al.* study in which female students outscored male students on participation and homework, but male students scored higher on exams for most semesters studied [74]. 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.

5.8 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. SEM showed that a latent variable called Conceptual Physics Performance/Non-Quantitative, CPP/NonQnt, 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.

Chapter 6

Gender Fairness of Physics Conceptual Inventories*

While the previous chapters explored differences in performance between male and female students, this chapter will explore intrinsic bias in the conceptual inventories themselves.

*This project 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.

The properties and performance of the most commonly used conceptual inventories constructed by physics education researchers have been studied through factor analysis [186–188], item response theory [189–192], and network analysis [193]. Most of these studies, however, have been performed using the FCI [21] and have only explored the structure and validity of undifferentiated samples. Substantially less research has been performed exploring the structure and validity of the FMCE [22], the CSEM [23], or the BEMA [24].

This chapter will examine the validity and fairness of the FCI, FMCE and CSEM using the validation framework proposed by Jorion *et al.* [142] for the evaluation of the validity of engineering conceptual inventories. The Jorion *et al.* framework begins with an examination of 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. IRT is then applied to further understand item functioning. Reliability is assessed with Cronbach alpha and inter-item correlations. Although not presented in this manuscript, factor analysis is then applied to understand sub-scale reliability.

This work will suggest an extension of the framework to include an item fairness analysis using Differential Item Functioning (DIF) analysis. The Educational Testing Service [194, 195], the American Educational Research Association (AERA), the American Psychological Association, and National Council on Measurement in Education [196] suggest that fairness analysis is a crucial step in instrument construction and, further, suggest DIF analysis as one part of fairness analysis. Item and instrument fairness is a sometimes contentious topic [197]. This work 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 on the material tested.

6.1 Gender Fairness within the Force Concept Inventory*

6.1.1 Introduction

Soon after the publication of the FCI, the reliability and validity of the instrument was questioned because the factor structure suggested by the authors was not recovered by exploratory factor analysis [186, 199, 200]. Exploration of the FCI factor structure has continued. In 2012, Scott, Schumayer, and Gray performed an exploratory factor analysis [187] which suggested a five-factor solution. Scott and Schumayer [192] supported this analysis on the same dataset with multitrait item response theory (MIRT).

CTT item difficulty and discrimination have been examined for the FCI. Wang and Bao identified three items to be problematic because they were too easy ($P > 0.8$) (items 1, 6, and 12) and two items were problematic because they were too challenging ($P < 0.2$) (items 17 and 26) [190]. Morris *et al.* also examined the FCI post-test items and found items 5, 17, and 26 were too challenging [201]. A study performed by Osborn Popp, Meltzer, and Megowan-Romanowicz was the only study to report item-level statistics for male and female students separately; post instruction, items 1, 6, and 16 were too easy for male students while item 26 was too challenging for female students [202].

The item-level properties of the FCI have also been examined using IRT [191, 202, 190, 192]. Wang and Bao explored the FCI at the item level using a three-parameter item response model (3PL model) [190]. In general, the 3PL IRT model fit the FCI data well. Planinic analyzed the FCI using the Rasch model and also found the FCI items functioned

*This manuscript is the first of two papers in Physical Review Physics Education Research. The research on this section has been published [198]. My role for the publication consisted of performing the various analyses and writing the results and discussion. I had primary responsibility for the production of this part of the work; however, the full publication was a collaborative effort with the co-authors.

properly [191].

The reliability of the FCI has been explored as well. Lasry *et al.* found the FCI demonstrated high internal consistency for both the initial application and the combination of the initial application and a retest that was given a week later [203]. Henderson also showed that the FCI has high test-retest reliability by examining the reliability between a graded post-test during the first quarter of introductory physics and an ungraded test given three weeks later during the second quarter of introductory physics [204].

Overall, while the structure and validity of the FCI is still a topic of active research, most studies have shown that most FCI items perform properly and can be well-modeled by IRT; however, most studies treat the sample as undifferentiated and do not examine male and female students separately.

6.1.2 Research Questions

- RQ1: Are there FCI items with difficulty, discrimination, or reliability values that would be identified as problematic within CTT or IRT? If so, are the problematic items consistent for male and female students?

- RQ2: Are there FCI items where the CTT or IRT difficulty is substantially different for male and female students?

- RQ3: Are there FCI items which DIF analysis identifies as substantially unfair to men or women?

- RQ4: Are unfair FCI items identified by item analysis?

- RQ5: Can differences in answering by men and women for problematic items be explained by an underlying physical principle or misconception?

- RQ6: If small to moderate and large effect DIF items are removed from the FCI, how does the gender gap change?

6.1.3 Methods

Samples

This study employs three datasets collected at four US universities. Racial/ethnic demographics were not available for individual students in the data and therefore, will be reported at the university level.

Sample 1: Sample 1 was collected from a large, southern land-grant university enrolling approximately 25,000 students. In 2012, university demographics by race/ethnicity were 79% white, 5% African American, 6% Hispanic, with other groups each 3% or less of the undergraduate population. It had a Carnegie classification of “Highest Research Activity” (or its precursor, “R1”) for the entire period studied. The range of ACT scores (25th percentile to 75th percentile) for the undergraduate population was 23-29 [173]. The sample was collected from the spring 2002 semester to the fall 2012 semester. The dataset contains 4,509 complete pretest responses (22.8% female) and 4,716 complete post-test responses (23.1% female).

The FCI was applied as a pretest and post-test in the introductory calculus-based mechanics class taken by scientists and engineers. Students received credit for a good faith effort on the pretest and received a grade on the post-test. The course was presented in the same format over the period studied and was overseen by the same lead instructor for all semesters studied. This instructor created all course materials including tests and homework assignments and was the lead lecturer for approximately 75% of the semesters studied. For the other semesters, a graduate student or visiting instructor familiar with the course delivered the lecture from the overall lead's notes. The course was presented with two 50-minute lectures and two 2-hour laboratory sessions each week. The lecture and laboratory components were tightly integrated. The lecture was traditional while the laboratory featured a combination of research-based methods including small -group problem solving, hands-on open or guided inquiry, and TA-led demonstrations, as well as traditional experiments. The course was revised to employ research-based techniques two years before the data collection for this study began. The revised course produced strong conceptual learning gains (Table 6.1) and was presented with few additional changes for the period studied. Because of the longitudinal stability of course oversight, content, and structure, this sample does not contain some of the confounding factors such as varying instructors bringing different coverage and class policy that might be present in other large datasets.

Sample 2: Sample 2 was drawn from two large urban public universities in the midwestern United States with similar student profiles (primarily regional commuter students with a moderate range of admission test scores). In 2014-2015, the first university in the sample had racial/ethnic demographics of 71% white, 13% African American, 7% international, with other groups 4% or less. The second university was 72% white, 10% African American, 6% Hispanic/Latino, with other groups 4% or less. The combined data contained 901 complete pretest responses (23.5% female) and 649 complete post-test responses (25.3% female). This sample includes data from fall 2014 to spring 2016 from several instructors.

Instructional styles ranged from traditional lecture, to moderately interactive lectures using Peer Instruction [25], to heavily interactive classes using Peer Instruction, Just-in-Time Teaching [205], and cooperative group problem-solving. Neither institution held a Carnegie classification of "Highest Research Activity" for the period studied. The range of ACT scores (25th percentile to 75th percentile) for one of the two institutions was 18-25 [173]. The other institution had a range of SAT scores (25th percentile to 75 percentile) of 890-1,130 which is equivalent to the 18-25 range of ACT scores [173].

Sample 3: Sample 3 was collected from a large eastern land-grant university enrolling approximately 30,000 students in the spring 2015 semester. In 2015, the university's racial/ethnic demographics for undergraduates were 81% white, 5% African American, 6% international, with all other categories 4% or less. Data collection was part of an effort to produce cross-norming data with an alternate mechanics conceptual evaluation routinely given at the institution and to explore the effects of distractor patterns on test performance [206]. Students received course credit for a good faith effort. Minor modifications (reordering the distractors) were applied to the FCI and found to have no significant effect. The FCI was applied to both the introductory, calculus-based mechanics and electricity and magnetism classes and therefore this sample contains a longitudinal component; the electricity and magnetism students had a larger time gap between instruction and testing than the mechanics students. The dataset contains 443 complete post-test responses (19% female); pretest data were not collected for Sample 3. This institution received the Carnegie classification of "Highest Research Activity" in the semester following the collection of the sample. The range of ACT scores (25th percentile to 75th percentile) for the undergraduate population was 21-26 [173].

The samples will be examined separately. The different post-test scores, instructional environments, and student populations (measured by ACT scores) did not suggest aggregating the samples would be productive. Further, because Sample 1 was much larger than Sample 2 and 3 combined, the aggregated dataset would largely produce the same results as Sample 1.

FCI Items

This and other studies have identified items which may be unfair either to men or women; a brief description of the most consistently identified items is provided. Item 6 is a Newton's 1st law problem about a ball after it has exited a circular track. Item 9 is a part of a group of items referring to a hockey puck sliding on a frictionless horizontal surface with a constant velocity. Item 9 asks about the speed of the puck just after it receives a kick. Item 12 asks about the trajectory of a cannon ball fired with initial velocity parallel to the ground. Item 14 asks about the trajectory of a bowling ball dropped from an airplane. Item 15 is a Newton's 3rd law problem involving a small car pushing a large truck. Items 21-24 are a group of questions about a rocket that is drifting sideways as its engine is turned on; the problems ask for the trajectory and change in speed with the engine on (21 and 22) and with the engine off (23 and 24). Item 27 asks how a box being pushed across the floor comes to a stop when the pushing force is removed.

Bonferroni correction

This work reports the statistical significance of many quantities and thus performs many statistical tests. A Bonferroni correction was applied to each set of analyses: $p < 0.05$

to $p < 0.0017$, $p < 0.01$ to $p < 0.00033$, and $p < 0.001$ to $p < 0.000033$ to correct for the 30 statistical tests performed for the 30 FCI items.

6.1.4 Results

		Male Students		Female Students		
	N	N	$(M \pm SD)\%$	N	$(M \pm SD)\%$	d
Sample 1						
Pretest	4509	3482	43 ± 18	1027	32 ± 14	0.69
Post-test	4716	3628	73 ± 17	1088	65 ± 18	0.46
Sample 2						
Pretest	882	673	43 ± 20	209	31 ± 15	0.66
Post-test	610	464	57 ± 24	146	45 ± 18	0.56
Sample 3						
Post-test	443	361	64 ± 20	82	51 ± 19	0.69

Table 6.1: Pretest and post-test averages for all samples. Mean M and standard deviation SD are reported as percentages. No pretest was given in the Sample 3 classes. Cohen's d measures the effect size of the difference between male and female scores.

Table 6.1 presents overall FCI pretest and post-test averages for the three samples. Significant gender differences (ps < 0.001) were measured for all applications of the FCI, with Cohen's d [130] indicating small to medium effect sizes. For Sample 1, course letter grades were available for about two-thirds of the participants. For this subset, female students ($M = 3.43$, $SD = 0.75$) had somewhat higher grades measured on a four-point scale than male students ($M = 3.24$, $SD = 0.89$) where M is the mean and SD the standard deviation. While there is substantial literature showing superior female performance on class grades [78] and superior male performance on standardized quantitative instruments [207, 86], this provides evidence that there was not a substantial disparity between male and female academic ability in Sample 1. The three samples present a spectrum of course outcomes with Sample 1 generating the highest scores on the FCI and Sample 2 the lowest. For Sample 1, female students closed the pretest gender gap of 11% somewhat to a post-test gap of 8%, while the gap changed little in Sample 2 from 12% on the pretest to 11% on the post-test.

Reliability and Correlation Analysis

As discussed in Chapter 3, Cronbach's alpha provides a measure of the overall reliability of an instrument. For Sample 1, the FCI was reliable with $\alpha = 0.84$ overall, male students $\alpha = 0.84$, and female students $\alpha = 0.83$. For male students, dropping item 29 increased alpha, while there was no item that could be removed to increase alpha for female students. For Sample 2, overall $\alpha = 0.90$ with $\alpha = 0.91$ for men and $\alpha = 0.81$ for women. For male and female students, there was no item whose removal increased alpha. For Sample 3, overall $\alpha = 0.86$: with $\alpha = 0.85$ for male students and $\alpha = 0.82$ for female students. Removing item 15 increased the overall alpha for both male and female students. These reliability values were consistent with those reported in Lasry *et al.* [203] and show that the FCI has strong internal consistency across a variety of instructional settings.

To further investigate reliability, the Pearson correlation coefficient between items was calculated. In general, if a student answers one item on a test correctly, the probability of answering a second item correctly should increase; item scores should be positively correlated. Jorion *et al.* [142] calculated tetrachoric correlations which assume the dichotomous variable, whether the question was correct or incorrect, was derived from an underlying normal continuum. This assumption seems unnatural for multiple-choice physics questions where the student must either answer completely correctly or incorrectly. Instead, the Pearson correlation was calculated for two dichotomous variables; the Pearson correlation is the ϕ

coefficient [140]. Tetrachoric correlations were also calculated and in all cases had absolute values greater than $|\phi|$. The significantly negatively correlated ($p < 0.05$) item pairs in Sample 1 were: male students, $\{23, 29\}$ and $\{29, 30\}$ and female students $\{8, 21\}$, $\{15, 27\}$, and $\{29, 30\}$. In Sample 2, there were no significantly negatively correlated item pairs for male students; for female students, only items $\{12, 29\}$ were significantly negatively correlated. For Sample 3, no question pairs were negatively correlated for men, while $\{7, 15\}$ and $\{9, 12\}$ were significantly negatively correlated for women. Both the correlation analysis and Cronbach's alpha support the identification of item 29 as problematic. Many of the items which were negatively correlated are identified as unfair in subsequent DIF analysis: items 9, 12, 15, 21, 23, and 27.

Difficulty and Discrimination

CTT and IRT were employed to examine the difficulty and discrimination of the FCI. Item-level post-test results for Sample 1 are presented in Table 6.3 and difficulty plotted in Fig. 6.1. The table presents the mean CTT difficulty P, CTT discrimination D, IRT difficulty b, and IRT discrimination a, for each FCI item. The CTT difficulties for Sample 2 and 3 are plotted in Fig. 6.2. Male and female students were investigated separately. The standard deviations for the CTT parameters were calculated by bootstrapping using 1000 sub-samples. Table 6.2 presents the problematic items identified in the FCI for each sample. Critically, many of the problematic items flagged for female students in Table 6.2 were not detected when the data remained aggregated over gender.

For Sample 1, all problematic items in the pretest had $P < 0.2$ (very hard) while all problematic post-test items had $P > 0.8$ (very easy). In Sample 2, all problematic pretest items had $P < 0.2$ while problematic post-test items for male students had $P > 0.8$ and problematic post-test items for female students had $P < 0.2$ (items 17 and 26) or $D < 0.2$ (item 29). For Sample 3, all problematic items had $P > 0.8$.

Examination of the gender-disaggregated post-test results in Table 6.2 identifies item 6 as problematic in 5 of the 6 samples while items 1, 12, and 29 were problematic in 4 of the 6 samples. Items 5, 17, 18, and 26 were problematic in all gender-disaggregated pretest samples. There was little additional commonality between the items flagged as problematic across all samples. The problematic items in the Sample 1 post-test all had very high scores. If the data was aggregated, item 12 was identified as problematic in all post-test samples.

Gender	Pre/Post	Problematic Items
Sample 1		
Female	Pre	5, 11, 13, 15, 17, 18, 25, 26, 28, 30
	Post	1, 3, 6, 7, 8, 9, 10, 12, 16, 19, 24, 29
Male	Pre	5, 6, 17, 18, 25, 26
	Post	1, 3, 6, 7, 8, 9, 10, 12, 13, 16, 19, 24, 29
Overall	Pre	5, 11, 17, 18, 25, 26
	Post	1, 3, 6, 7, 8, 9, 10, 12, 13, 16, 19, 24, 29
Sample 2		
Female	Pre	2, 5, 11, 13, 17, 18, 20, 25, 26, 28, 30
	Post	17, 26, 29
Male	Pre	5, 17, 18, 26
	Post	6, 12
Overall	Pre	5, 11, 13, 17, 18, 26
	Post	12
Sample 3		
Female	Post	1, 4, 6, 29
Male	Post	1, 4, 6, 7, 12, 16, 24
Overall	Post	1, 4, 6, 7, 12, 16, 24

Table 6.2: CTT problematic items with $P < 0.2$, $P > 0.8$, or $D < 0.2$.

IRT results can also be used to identify problematic items. One FCI item, item 29, produced difficulty parameters indicating the IRT model was a poor fit for that item. None of the FCI items showed the dramatic departures from model fit including negative discrimination parameters identified in some of the inventories examined by Jorion *et al.* [142]. As such, IRT supports the identification of item 29 as problematic.

	Classical Test Theory					Item Response Theory							DIF	
#	P_M	P_F	D_M	D_F	ϕ	b_M	b_F	a_M	a_F	d	V_M	V_F	$\Delta\alpha_{MH}$	L
1	.97±.00	.95±.01	.10±.01	.13±.02	.04	-2.71±.16	-2.78±.32	1.63±.14	1.30±.21	.01	.02	.03	.33	.10
2	.66±.01	.60±.02	.56±.02	.44±.04	$.05^b$	-.74± .05	-.61±.11	1.09±.06	.73±.08	.05	.02	.03	.44	.50
3	.91±.00	.90±.01	.22±.01	.25±.03	.01	-2.15±.10	-1.77±.12	1.42±.09	1.89±.20	.07	.02	.04	1.17^b	.84
4	.62±.01	.62±.01	.59±.02	.57±.03	.00	-.57±.04	-.54±.07	1.05±.05	1.19±.11	.01	.02	.05	1.28^c	1.26^c
5	.58±.01	.50±.01	.63±.02	.65±.03	$.07^c$	-.35±.04	-.03±.07	1.24±.06	1.08±.10	$.15^c$	$.02^a$	$.06^a$	.50	.35
6	.91±.00	.80±.01	.22±.01	.34±.03	$.15^c$	-2.34±.12	-2.07±.25	1.23±.09	.75±.10	.04	.02	.03	-1.43^c	-1.34^c
7	.88±.01	.81±.01	.22±.02	.28±.03	$.08^c$	-2.69±.19	-2.64±.40	.81±.07	.58±.10	.00	.02	.03	-.45	-.21
8	.89±.01	.84±.01	.26±.01	.37±.03	$.08^c$	-2.13±.11	-1.44±.10	1.26±.08	1.62±.16	$.12^c$	.02	$.06^a$	-.14	-.17
9	.80±.01	.84±.01	.38±.02	.40±.03	.03	-1.56±.08	-1.44±.10	1.12±.06	1.59±.15	.03	$.03^c$	$.06^a$	1.89^c	1.76^c
10	.93±.00	.90±.01	.21±.01	.28±.03	$.05^a$	-1.99±.08	-1.72±.11	1.95±.13	1.92±.21	.06	.02	.05	.39	.08
11	.76±.01	.73±.01	.53±.02	.63±.03	.02	-1.05±.04	-.82±.06	1.53±.07	2.15±.18	$.09^a$	$.03^c$	$.06^b$	1.31^c	$.87^b$
12	.93±.00	.80±.01	.16±.01	.31±.03	$.17^c$	-3.06±.22	-2.16±.27	.94±.08	.71±.10	.07	.02	.02	-1.97^c	-1.84^c
13	.83±.01	.79±.01	.50±.02	.57±.03	.04	-1.16±.04	-.99±.06	2.39±.12	2.51±.22	.08	.02	$.05^a$	1.22^c	.53
14	.67±.01	.40±.01	.46±.02	.44±.04	$.23^c$	-1.01±.07	.63±.12	.78±.05	.66±.08	$.39^c$	$.02^a$	$.06^b$	-1.97^c	-1.84^c
15	.60±.01	.66±.02	.45±.02	.54±.04	$.05^b$	-.64±.06	-.71±.07	.72±.05	1.28±.11	.02	$.05^c$	$.08^c$	1.77^c	2.00^c
16	.94±.00	.91±.01	.17±.01	.28±.03	.04	-2.33±.11	-1.71±.11	1.51±.11	2.15±.24	$.10^b$	.02	.05	.36	.17
17	.55±.01	.49±.02	.67±.02	.62±.03	$.05^a$	-.19±.03	.03±.07	1.42±.06	1.19±.10	$.11^a$	.02	.05	$.84^c$	.62
18	.57±.01	.52±.02	.68±.02	.69±.03	.04	-.27±.03	-.09±.06	1.44±.06	1.27±.11	.09	.02	.05	1.04^c	$.70^a$
19	.87±.01	.87±.01	.29±.02	.33±.03	.00	-1.86±.09	-1.65±.12	1.28±.08	1.56±.16	.04	.02	$.06^b$	1.35^c	1.14^c
20	.65±.01	.61±.01	.53±.02	.55±.03	.03	-.74±.05	-.57±.09	1.00±.05	.95±.09	.06	.02	.04	$.75^b$	$.77^b$
21	.47±.01	.23±.01	.60±.02	.29±.04	$.20^c$	.14±.04	2.25±.33	.99±.05	.57±.08	$.38^c$	$.04^c$	.05	-1.86^c	-1.77^c
22	.58±.01	.34±.01	.60±.02	.42±.04	$.20^c$	-.38±.04	1.11±.16	1.08±.05	.64±.08	$.45^c$	$.03^c$	$.07^c$	-1.61^c	-1.56^c
23	.77±.01	.45±.02	.45±.02	.43±.04	$.29^c$	-1.31±.06	.35±.13	1.15±.06	.55±.08	$.43^c$	.02	.03	-2.70^c	-2.71^c
24	.92±.00	.83±.01	.20±.01	.32±.03	$.12^c$	-2.38±.13	-1.79±.16	1.26±.09	1.10±.12	.08	.02	.04	$-.94^b$	$-.98^b$
25	.54±.01	.46±.01	.74±.02	.66±.03	$.07^c$	-.17±.03	.14±.06	1.72±.08	1.31±.11	$.17^c$	$.03^c$	$.06^a$	$.70^a$	.32
26	.32±.01	.23±.01	.66±.02	.51±.04	$.09^c$	.64±.04	1.15±.09	1.65±.08	1.44±.13	$.22^c$	$.03^c$	.05	.40	-.08
27	.77±.01	.53±.02	.38±.02	.37±.04	$.22^c$	-1.58±.09	-.27±.15	.86±.05	.45±.07	$.24^c$	$.02^a$	$.06^a$	-1.87^c	-1.80^c
28	.71±.01	.66±.01	.63±.02	.62±.03	$.05^a$	-.83±.04	-.65±.07	1.50±.07	1.37±.12	.08	$.02^a$	.05	$.83^b$	.56
29	.83±.01	.85±.01	.09±.02	.14±.03	.02	-18.4±10	-5.24±1.47	.09±.05	.34±.10	.02	$.03^c$	.04	.64	1.55^c
30	.62±.01	.53±.01	.59±.02	.55±.04	$.08^c$	-.52±.04	-.16±.08	1.24±.06	.86±.09	$.15^b$	.02	.03	.19	.18

Table 6.3: Classical Test Theory and Item Response Theory results for Sample 1 for each FCI item. Male results are marked M and female results F. Significance levels have been Bonferroni corrected for the number of statistics tests: "a" denotes $p < 0.0017$, "b" $p < 0.00033$, and "c" $p < 0.000033$.

Item Fairness

An item is "fair" if students of the same ability from two populations produce equal scores on the item. We first investigate item fairness under the assumption that male and female students are of equal abilities, then apply DIF analysis to explore fairness without the assumption of equal abilities. For this analysis, Samples 2 and 3 contain an insufficient number of female students to draw strong statistical conclusions. The results for these samples are examined only in reference to Sample 1.

This work uses the terms "ability" and "fairness," which are common within the test development literature [194]. Both terms have broad colloquial meanings outside this literature, and as such, it is important that the reader interpret these terms by their narrow meaning. Ability is used to mean only the proficiency with which students answer test items—in this case, conceptual physics problems on the FCI. Fairness analysis depends on the assumptions made about ability. If two groups have the same proficiency in conceptual physics, then items where the groups score differently do not test the two groups in the same way: the items are unfair. If the assumption of equal proficiency is not true, then items can score differently because of the differences in the groups and a difference in score does not imply an unfair problem. DIF analysis does not assume the two groups have equal proficiency in conceptual physics, but uses the score on the FCI as a measure of proficiency. In DIF analysis, an item is unfair if the two groups have a larger difference in score than one would predict from the difference in overall test score.

DIF analysis uses the overall test score as a measure of ability and, therefore, cannot detect if items in an instrument are generally unfair. It can only detect when an item is

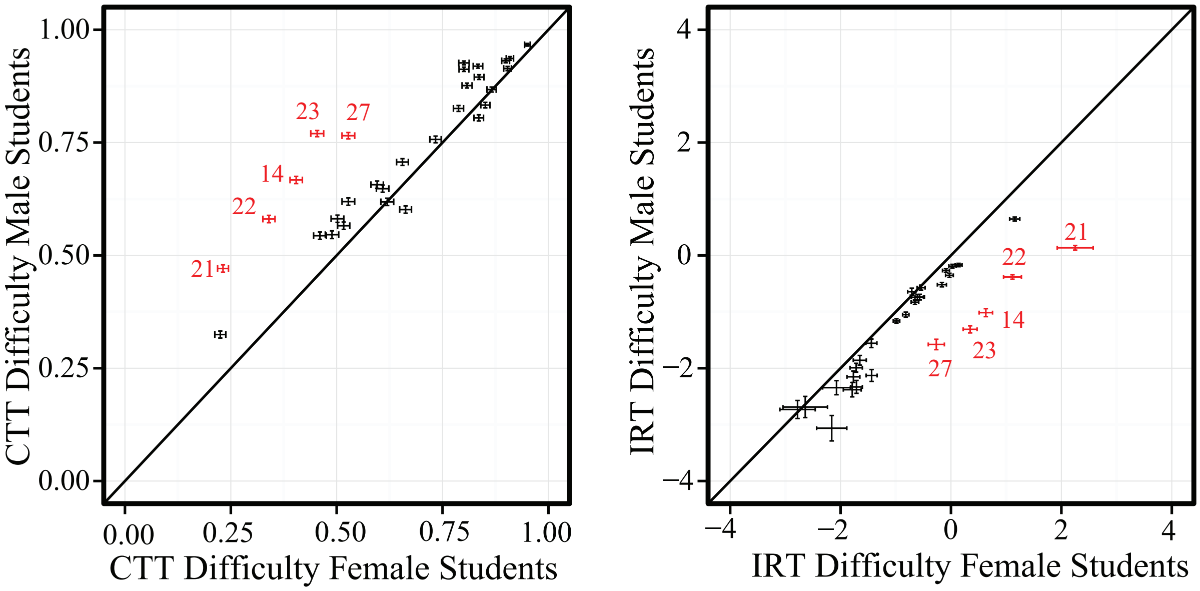

Figure 6.1: CTT and IRT post-test results for Sample 1. Items 14, 21, 22, 23, and 27 are marked in red and labeled. A line of slope one is drawn to allow comparison of male and female difficulty. Error bars represent one standard deviation in each direction.

functioning differently than the overall instrument.

Graphical Analysis If one assumes that male and female students have an equal ability to answer conceptual physics questions correctly, then a fair FCI item is one where the difficulty is equal for male and female students. Under this assumption, which is supported by the higher course grades of female students, item fairness can be explored by plotting the difficulty for male students against the difficulty for female students. Figure 6.1 shows this plot for the Sample 1 post-test. A line of slope one is drawn on all plots; perfectly fair questions would fall on this line (the fairness line). Items unfair to women fall above the fairness line for the CTT plots and below the line for IRT plots. Fig. 6.1 has three striking features: (1) most items are significantly unfair to women (the error bars do not overlap the fairness line); (2) five items, 14, 21, 22, 23, and 27, stand out as substantially unfair to women by falling well off the fairness line; and (3) most other items fell fairly close, but on the unfair to women side, of the post-test fairness line. The substantially unfair items are plotted in red and numbered in the figure. Similar plots were explored for item discrimination and did not show any pattern of item bias; therefore, the remainder of the study will focus on item difficulty.

To determine if the differences in performance in the CTT plot in Fig. 6.1 were statistically significant and to estimate effect sizes, the ϕ coefficient was calculated for each item and is included in Table 6.3. The significance values for ϕ were calculated using a χ^2 test of independence between male and female correct and incorrect answers for each item. For many items, male and female scores were significantly different. For items 6, 12, 14, 21, 22, 23, 24 and 27, male and female difficulty scores were significantly different with a small effect size. This set of items contains most of the items which will be identified as significantly unfair by DIF analysis.

The ϕ coefficient above is mathematically similar to the ϕ coefficient described previously in the reliability and correlation analysis; however, their use is conceptually different. Above, ϕ is used as a measure of independence and large ϕ indicates that the item difficulty is different for men and women (small ϕ indicates the difficulty is independent of gender). In the previous reliability and correlation analysis, ϕ is used as a measure of association, so large ϕ indicates strongly correlated items.

A similar analysis was used to explore whether differences in the IRT difficulty coefficients were significant. The differences are characterized by Cohen's d (listed in Table 6.3). The results were similar to those using the CTT difficulty; the gender difference in items 14, 21, 22, 23, 26 and 27 was significant (ps< 0.001) with a small to medium effect size. Table 6.3 also presents measures of the goodness-of-fit of the IRT model for men and women

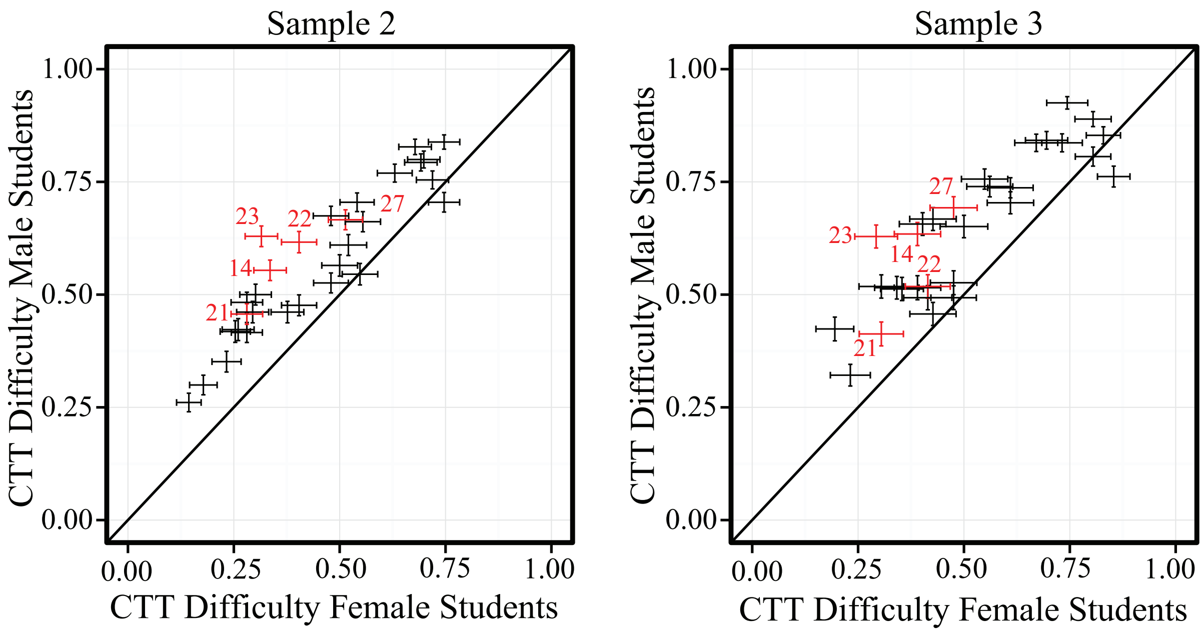

Figure 6.2: CTT post-test difficulty results for Sample 2 and 3. Items 14, 21, 22, 23, and 27 are marked in red. A line of slope one is drawn to allow comparison of male and female difficulty.

through Cramer's V statistic.

One item, item 29, produced difficulty and discrimination parameters that suggest the underlying IRT model was a poor approximation for this item. The model was re-fit removing this item. Parameter estimates changed very little; as such, the values for the original model including item 29 are reported.

Figure 6.2 presents a plot of CTT post-test difficulty for Samples 2 and 3 with items 14, 21, 22, 23, and 27 also colored in red and labeled. The much smaller sample size caused the error bars of many points to overlap, but many of the five most problematic items in Sample 1 were also at the outside of the item envelope in Samples 2 and 3.

Figure 6.3 overlays plots of items 14, 21, 22, 23, and 27 for all samples; the similarities, particularly in the CTT plot, are quite strong. This supports the identification of these five questions as generally unfair, not simply unfair because of some artifact of either student

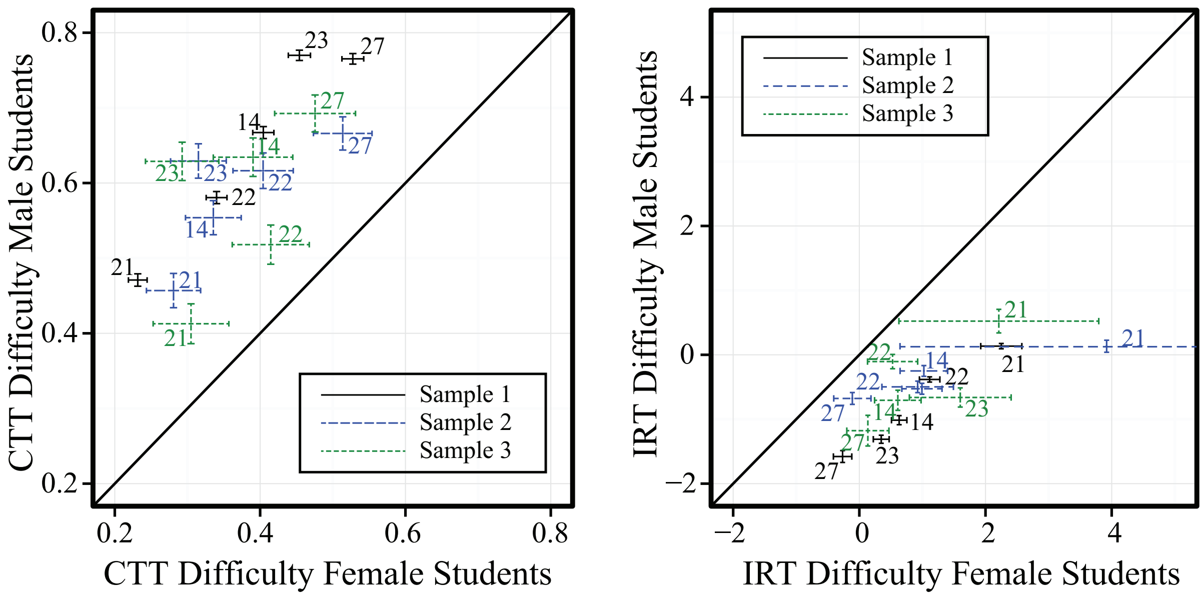

Figure 6.3: CTT and IRT post-test difficulty scores for male and female students for problematic items from all samples. A line of slope one is drawn to allow comparison of male and female difficulty. The item number for each problem is also labeled. The IRT difficulty of Sample 2, item 23 is not labeled; the point overlays that of Sample 2, item 22.

population or instruction in Sample 1. IRT results for Samples 2 and 3 are included in Fig. 6.3, but should be interpreted with caution, as these samples were too small for reliable IRT parameter estimation.

Differential Item Functioning Analysis The analysis of the previous section compared male and female students and found significant differences in difficulty for many FCI items under the assumption of equal male and female ability. The clustering of many items near the fairness line in Fig. 6.1 suggests that, while there may be some overall difference in conceptual performance between men and women, most items were only somewhat more difficult for women than men.

DIF analysis relaxes the assumption of equal ability and replaces it with the assumption that the overall score on the instrument is an accurate measure of ability. Table 6.3 reports

$\Delta\alpha_{MH}$ for each item in Sample 1. Eight FCI items demonstrated large DIF (9, 12, 14, 15, 21, 22, 23, 27), where 9 and 15 were biased in favor of female students. This set includes most items identified as significantly unfair with a small effect size in the previous section. Seven additional questions demonstrated small to moderate DIF.

DIF analysis can also be carried out using the results of IRT. We used Lord's statistic L, which is mapped to the same range as $\Delta\alpha_{MH}$ and reported in Table 6.3. The Lord's statistic results agreed with the high DIF classification provided by $\Delta\alpha_{MH}$ except that item 29 was also flagged as high DIF favoring women. The small to moderate DIF results were less consistent, and the two statistics disagreed on items 3, 11, 13, and 18. None of these four items were ultimately identified as biased in the reduced FCI instrument constructed to answer RQ6. This provides evidence of the efficacy of employing both CTT and IRT analysis to complement one another.

DIF analysis was also attempted for Samples 2 and 3 by stratifying students into five quantiles to reproduce the analysis of Dietz *et al.* [117]. The stratification into 5 quantiles left only a few women in the highest scoring quantile and the results were strongly dependent on the number of quantiles selected. We concluded that the number of female students in Samples 2 and 3 was insufficient for accurate DIF analysis.

Item-level Analysis

The distribution of student answers for the five most unfair items of Sample 1 are shown in Table 6.4. Female students preferentially selected one of the distractors for each item. For Samples 2 and 3, the selection of distractors was less uniform, possibly because of the relatively small number of female students in Samples 2 and 3 or because of the lower overall FCI scores for these samples. The differences in responses observed between male and female students in Sample 1 may have resulted from one or more physics concepts that were not mastered by female students or from surface features of the problem's context that made the problem more difficult for female students. Examination of these problems does not immediately suggest a common physics concept underlying the incorrect answers.

For item 14 (bowling ball falling out of an airplane), the most popular distractor for female students was the rearward parabolic trajectory, while the most popular distractor for male students was a linear forward trajectory. Item group 21–24 concerns a scenario where a sideways-drifting rocket turns on its engine for a period and then off again. The differences in items 21 to 23 seemed to result from students answering the question correctly for the assumption that the force was an impulse force. The preferentially selected distractor for items 21 and 22, for both men and women, was correct for an impulse force. The relatively random pattern of incorrect answers on item 23 (turning off the engine) might result because the question does not make sense if one is assuming the engine is already off. The question group does state that the engine is on for the entirety of items 21 and 22. The text employs the verb "thrust"; colloquially, the verb "to thrust" means to "push or drive quickly and forcibly" [208]. Item 27 concerns a large box being pushed across a horizontal floor, and the preferred distractor

		Response				
#	Gender	(a)	(b)	(c)	(d)	(e)
14	Male	10%	4%	18%	**67%**	0%
	Female	30%	12%	17%	**40%**	0%
21	Male	2%	5%	39%	7%	**47%**
	Female	3%	16%	53%	5%	**23%**
22	Male	31%	**58%**	1%	9%	0%
	Female	55%	**34%**	1%	9%	0%
23	Male	7%	**77%**	6%	8%	1%
	Female	25%	**45%**	13%	14%	2%
27	Male	19%	3%	**77%**	1%	0%
	Female	40%	6%	**53%**	1%	0%

Table 6.4: Answer distribution for problems with large gender differences in CTT and IRT difficulty in Sample 1. Correct answers are bolded.

across genders was that the box comes immediately to a stop.

The problem contexts described above might be more familiar on average to men through everyday experience (item 27) or through greater exposure to physically realistic video games and movies (items 14, 21–23). Wilson *et al.* showed that gender differences in physics questions used in physics competitions were particularly large for two-dimensional motion and projectile motion problems [120]. However, questions identified in the current study as unfair to both men and women fall in these categories. Without the identification of a physical principle or common misconception that unifies the items, the determination of the origin of the gender difference must be left for a future study.

An Unbiased Force Concept Inventory

To construct an unbiased version of the FCI, items were iteratively removed, $\Delta\alpha_{MH}$ recalculated, and additional items removed until no item in the FCI showed small to moderate or large DIF for Sample 1. This process removed the 8 questions with large DIF as well as items 6 and 24, producing a reduced instrument containing FCI questions: 1, 2, 3, 4, 5, 7, 8, 10, 11, 13, 16, 17, 18, 19, 20, 25, 26, 28, 29, and 30. For Sample 1, this 20-item instrument reduced the gender gap on the post-test to 4.3% from the original 8.0%, with men scoring (73.1±19)% and women scoring (68.7±19)%. The difference was still significant $[t(1761) = 6.55, p < 0.001]$ but with a substantially smaller effect size, $d = 0.23$. The total scores on the original and reduced instruments were highly correlated for both male and female students (Pearson correlation $r = 0.96$). If the instrument is further reduced by removing item 29, which was flagged by item analysis and by Lord's statistic, the gender gap increases slightly to 4.7%. The reduced instrument still contains a number of items originally calculated to have small to moderate DIF (see Table 6.3). The DIF of these items became negligible after the higher DIF items were removed.

For Samples 2 and 3, the reduced instrument did not substantially reduce the gender gap. For Sample 2, the original gender gap of 12.9% became 11.4% for the 20-item instrument and 12.2% with the further removal of item 29. For Sample 3, the original gender gap of 13.5% was reduced to 12.7% for the 20-item instrument, but increased to 13.8% with the removal of item 29.

Pretest Results

The FCI pretest was analyzed using the same methods as the post-test. Fig. 6.4 shows the FCI pretest results for Sample 1. The five substantially unfair questions identified in the post-test (14, 21, 22, 23, 27) were among the most unfair questions in the pretest plots. However, many additional questions were also substantially more difficult for women. The IRT variance for women was also substantially higher than in the post-test. Many pretest differences were reduced by instruction and many questions moved substantially closer to the fairness line in the post-test, except items 14, 21, 22, 23, and 27.

DIF analysis was also performed on the Sample 1 pretest. With the much larger variance seen in Fig. 6.4 and the generally weaker pretest performance of women, few items were detected as significantly biased. The DIF results for the FCI pretest for Sample 1 detected only item 14 as having large DIF; items 4, 12, 19, and 26 demonstrated small to moderate DIF. This difference between pretest and post-test is consistent with the observation that women close the score gap with men on many problems post instruction. Because DIF stratifies by overall test score, a smaller gap can be considered unfair on the post-test than the

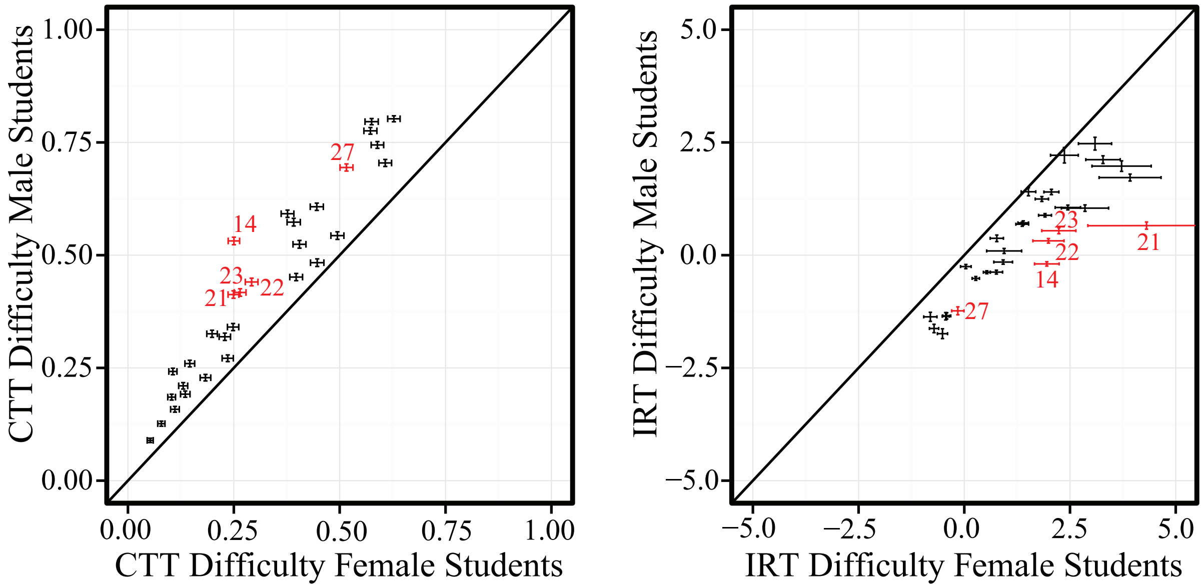

Figure 6.4: CTT and IRT pretest results for Sample 1. Items 14, 21, 22, 23, and 27 are marked in red and labeled. A line of slope one is drawn to allow comparison of male and female difficulty. Error bars represent one standard deviation in each direction.

pretest if the overall post-test gap is smaller than the pretest gap.

This study found a 20-item unbiased version of the FCI. The pretest gender gaps changed little on the reduced unbiased instruments. For Sample 1, the pretest gender gap on the 20-item FCI was 9.9% which was somewhat smaller than the gender gap of 11.9% on the original 30-item FCI. Further removing item 29 increased the gap to 10.1%. For Sample 2, the gender gap on the 20-item instrument was 10.3% which was somewhat smaller than the gender gap of 12.3% on the original 30-item FCI. Further removing item 29 increased the gap to 10.6%.

6.1.5 Discussion

This study sought to answer six research questions; these are addressed in the order proposed.

RQ1: Are there FCI items with difficulty, discrimination, or reliability values that would be identified as problematic within CTT or IRT? If so, are the problematic items consistent for male and female students? CTT identified few areas where the FCI or items within the FCI were uniformly problematic across all samples. Aggregating men and women, item 12 was flagged as problematic in all post-test samples. Items 5, 11, 17, 18, and 26 were identified as problematic in both aggregated pretest samples. Item 6 was problematic in 5 of the 6 gender-disaggregated post-test samples. Items 1, 12, and 29 were identified as problematic in 4 of the 6 gender-disaggregated post-test samples. Items 5, 17, 18, and 26 were identified as problematic in all gender-disaggregated pretest samples. Identification of difficulty parameters outside the desired range likely resulted from the application of the FCI at multiple institutions with differing student populations as both a pretest and post-test. This caused some items to be flagged on the pretest with $P < 0.2$ and on the post-test with $P > 0.8$. IRT and reliability analyses further supported the identification of item 29 as problematic.

The items and the number of items identified as problematic differed between male and female students. More items were problematic for female students in Sample 1 and Sample 2 on the pretest. More items were problematic for male students in Sample 3 on the post-test. Crucially, an analysis that aggregated men and women, the "Overall" rows in Table 6.2, would reach conclusions accurate for male students but often very inaccurate for female students.

The problematic CTT and IRT items provide less accurate information about the knowledge of the student than non-problematic items by either being too hard, too easy, or too likely to answered correctly by weak students (or incorrectly by strong students). Many items on the FCI provide less information about female students than male students in the Sample 1 and 2 pretest; the FCI contains many items that provide less information about male students in the Sample 3 post-test. While these problems almost certainly resulted from using one instrument in multiple environments both as a pretest and post-test, instructors should be aware that the FCI can provide results with different levels of validity for different student populations even in the same testing conditions. As such, its results should used with caution for these populations.

RQ2: Are there FCI post-test items where the difficulty is substantially different for male and female students? FCI items 6, 12, 14, 21, 22, 23, 24, and 27 in Sample 1 demonstrated a significant gender bias in item difficulty (Table 6.3) in CTT with a small effect size. IRT identified items 14, 21, 22, 23, 26, and 27 as significantly unfair with a small effect size. The interpretation of items 14, 21, 22, 23, and 27 as substantially unfair was supported by graphical analysis of Samples 2 and 3 (Fig. 6.3).

RQ3: Are there FCI items which DIF analysis identifies as substantially unfair to men or women? In Sample 1, DIF analysis confirmed the unfairness of items 12, 14, 21, 22, 23, and 27 and further identified items 9 and 15 as having large DIF; items 9 and 15 were biased in favor of women. Iteratively removing high DIF items also showed items 6 and 24 with high DIF once the highly biased items were removed. Because DIF depends on overall test score, the DIF of an item changes as unfairly functioning items are removed from an instrument. Items 3, 4, 11, and 18 demonstrated small to moderate DIF; however, the DIF of these items became negligible as the more unfair items were removed to form the 20-item unbiased FCI.

The Sample 1 post-test results of this study were fairly consistent with those of other work. The Sample 1 results of this study supported the advantage for women in item 9 found in Deitz *et al.* [117] (large DIF) and Osborn Popp *et al.* [202] (small to moderate DIF). This study also supported the large DIF toward men of item 23 found in both of these previous studies. Deitz *et al.* did not report small to moderate DIF items; however, from the graph presented [117, Fig. 4] it seems likely item 15 would be found biased towards women and items 12, 14, and 27 biased towards men, consistent with this work. The graph also suggests item 30 may also be biased toward men. Osborn Popp *et al.* also identified items 4, 9, 15, and 29 with small to moderate DIF toward women and items 6 and 14 with small to moderate DIF toward men. The current study identified item 4 as unfair (small to moderate DIF) in Sample 1, as was reported in Deitz *et al.* (large DIF) and Osborn Popp *et al.* (small to moderate DIF); however, the DIF of this item became negligible as more highly biased items were removed from the FCI. Items 14, 22, 23, and 29 were also identified by McCullough and Meltzer as demonstrating significant differences between male and female answering patterns when the context of the question was modified to be more stereotypically female oriented [116].

Combining the results of this study with those of previous research strongly identifies a set of unfair items in the FCI. The relatively consistent pattern of items 6, 9, 12, 14, 15, 22, 23, and 27 being identified as gender biased in multiple studies strongly indicates the use of these questions should be reconsidered. This study additionally suggests that items 21 and 24 should be reconsidered because of bias and item 29 because of recurring reliability issues. Removing all these items would produce a 19-item instrument. Because the FCI has not demonstrated a consistent factor structure [186] and therefore is primarily a single factor instrument measuring the degree to which a student possesses a "Newtonian Force

Concept," a 19-item instrument should measure this construct with approximately the same accuracy as a 30-item instrument.

RQ4: Are unfair FCI items identified by item analysis? Most items ultimately identified as unfair in the FCI were not uniformly flagged as problematic by CTT or IRT item analysis. Only items 6 and 12 were detected as problematic in both DIF and item analysis using discrimination and difficulty cutoffs. Item fairness analysis is therefore a complementary method that provides additional information beyond item analysis methods. CTT and IRT difficulty, discrimination, and reliability checks do not guarantee item score fairness.

RQ5: Can differences in answering by men and women for problematic items be explained by an underlying physical principle or misconception? Examining answer patterns for the biased questions in Sample 1 did not identify an underlying physical principle or misconception that was shared by all or some combination of the questions. This makes it unlikely a general failure of instruction either by the course studied or within the academic background of the students studied accounted for the differences identified. Further experimental investigation such as that performed by McCullough and Meltzer [116] will be required to determine the origin of the gender differences.

RQ6: If small to moderate and large effect DIF items are removed from the FCI, how does the gender gap change? For Sample 1, removal of all questions with small to large DIF resulted in a 20-item instrument. The gender gap on the post-test using this reduced instrument was 4.3% ($d = 0.23$) which was substantially smaller than the original post-test gender gap of 8.0% ($d = 0.46$) with half the effect size. Item fairness, then, does not explain all the gender gap in the FCI but accounts for about half of the gap in this sample. The gender gap on the 20-item gender-neutral instrument's post-test would be the second smallest FCI gap reported [63].

The reduced instrument did not significantly reduce the gender gap in Samples 2 and 3. An explanation to this result may be found by comparing Fig. 6.1, Fig. 6.2, and Fig. 6.4. In Sample 1, female students improved on many items that were substantially unfair in the pretest, leaving only a few items where women were substantially off the fairness line on the post-test. Sample 2 and 3 students did not demonstrate the same degree of progress, and women in these samples do not show a substantial number of nearly fair questions post-instruction.

Some studies have suggested that more interactive teaching methods lower the gender gap [72, 73, 122]; however, this effect has not been consistently reproduced [124–126]. Some research-based instructional methods were employed in the lecture portions of Sample 2 and 3 while Sample 1 combined a traditional lecture with an interactive, inquiry-based laboratory experience. While the courses from which all three samples were drawn presented some interactive or research-based instruction, the primary differences between the courses seems to be the overall conceptual learning outcome measured by FCI post-test scores. Excluding the items showing substantial gender bias, the course measured in Sample 1 produced post-test results where the performance of male and female students were more similar (most results fell near the fairness line). The post-test results for Sample 2 and 3 have many more items substantially off the fairness line. Examination of the Sample 1 pretest plots showed many more items substantially off the fairness line; the instruction in the class moved female students nearer the fairness line on many items (except the gender biased items). This comparison suggests that it is not only the interactivity of the instruction that matters in reducing the gender gap but also its overall effectiveness. It seems possible that the gender gap closes for interactive courses only if they produce superior learning outcomes, measured by FCI post-test scores. This could explain the inconsistent relationship between interactive instruction and lowering the gender gap [122, 72, 73, 124–126].

Comparing results for Samples 1, 2, and 3 illuminates the variability of previous research into item fairness. While not as large as Sample 1, Samples 2 and 3 contain as many or more students than some of the other studies of item fairness. Difficulty measures for these samples had large error bars, particularly for female students. Both samples also involved confounding factors such as multiple instructors and pedagogies or a longitudinal application of the FCI which would also increase variability. The gender biased items were hidden by the noise in these samples and were probably partially obscured by variation in other studies. Experiments sub-sampling Sample 1 suggest 1000–1500 as a minimum sample size to clearly resolve gender disparities in FCI datasets where women are significantly underrepresented.

The inclusion of many unfair items calls into question the practical application of the FCI instrument as well as research based on the FCI. Examples of the threat to research validity can be found in two recent studies. In a factor analysis of the FCI [187], gender biased items 21, 22, 23, and 27 factored together while item 14 failed to be included in any factor. This raises the question of whether the gender bias of the questions influenced the factor structure.

Han *et al.* [209] investigated dividing the FCI into two shorter tests (half-tests) to lower the time burdens of testing. Gender fairness was not considered in their analysis. Randomly, four of the five highly unfair to women questions (14, 21, 22, and 23) were included in the second half-test while none of the highly unfair questions were included in the first. The second half-test also included item 24 which was identified as unfair after highly unfair items were removed from the FCI. The first half-test also contained the two questions that DIF identified as biased toward women (9 and 15) and two of the additional questions DIF identified as biased toward men (6 and 12). As such, it is likely that the second half-test is more gender unfair than the FCI and the first half-test is more gender neutral.

This study identified a reliable and fair 19-item version of the FCI. It seems likely, however, that if this instrument were deployed in diverse educational settings as both a pretest and post-test that it would produce results with differing levels of validity for men and women in some situations by posing questions that are either too easy or too hard for the student population. As such, instructors using this instrument should be aware of the possibility of unfairness and either confirm the fairness of the instrument independently or restrict the kinds of decisions made from the results of the instrument. For example, using the FCI pretest as a baseline measurement without instructional consequences may be appropriate, but using pretest scores to assign lab groups may not be.

6.1.6 Implications

This work identified multiple questions within the FCI which were unfair to either men and women; this finding was supported by multiple samples and is consistent with other studies reporting unfair items. As such, we suggest the use of the score on the full 30-item FCI be discontinued and the 19-item unbiased score used in the future. Institutions with longitudinal FCI datasets should convert FCI scores to the 19-item unbiased scoring. The full 30-item score should continue to be reported to allow comparison with previous research.

6.1.7 Limitations

While this research used data from four institutions combined to form three datasets, two of the datasets were too small to provide adequate statistical power to determine if some conclusions were general. The analysis should be conducted with additional large datasets to determine whether the conclusions are widely replicated. In addition, the fairness analyses were not sensitive to differences in slopes. An alternative approach would be to use muli-group IRT to test for both difficulty and discrimination.

6.1.8 Conclusions

The FCI is broadly used to assess physics instruction and conceptual learning. The above analysis demonstrated that it contains a number of items that are not fair to women and a few items unfair to men. The prevalence of the FCI and large longitudinal datasets that have been collected make it difficult to suggest that its use should be discontinued; however, the 30-item score should not be used for any purpose from which a student might benefit. We suggest the continued reporting of the full FCI score along with the score on the reduced unbiased instrument. The reduced unbiased instrument score should be used for instructional decisions and to assign course credit.

The reporting of gender composition is uneven in PER. Researchers referencing FCI scores at multiple institutions should be aware that these scores may contain variation that results from gender differences that were not reported.

By most measures available to conceptual inventory developers where limited initial deployment is possible, the FCI performs exceptionally well. The identification of the unfair items required multiple studies and very large samples. As such, future developers of conceptual instruments should plan for a second level of validation which can only be carried out if their instrument achieves broad deployment. This validation might identify items with unexpected biases, reliability, or validity problems. The overall instrument and any sub-scales should be sufficiently robust that the removal of some items leaves the validity and reliability of the instrument intact.

6.2 Gender Fairness within the Force and Motion Conceptual Evaluation*

6.2.1 Introduction

While there is a substantial body of research investigating the gender gap of the FCI, little research into the gender fairness of the FMCE exists. Although most of the research on the FMCE examines overall scores pre- and post-instruction, 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 [210]. 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 [211]. 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" [212]. For example, the three

*This section is a portion of the second of the two-part manuscript in Physical Review Physics Education Research. This manuscript is currently under review. I am the primary author of this work; however, the final product was produced in collaboration with the co-authors.

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.

The three distinct clusters described by Thornton *et al.* have been studied at the item level [213, 214]. 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* [213]. In 2014, Smith, Wittmann, and Carter used these revised clusters to provide insight into the effectiveness of interactive classroom techniques [214].

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 [215]. To our knowledge, little research has been performed investigating the reliability, validity, or fairness of the FMCE.

Overall, the analyses of the FMCE at the item-level treat the students as an undifferentiated sample; the comparison of CTT difficulty and discrimination between male and female students for the FMCE has not yet been reported.

6.2.2 Research Questions

- RQ1: Are there items in the FMCE which CTT would identify as problematic? Are the problematic items the same for male and female students?

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

6.2.3 Methods

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 [216]. The general undergraduate population had a range of ACT scores from 25-30 (25th to 75th percentile range) [173]. 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% [173].

The course was taught by four faculty members and shared a common format throughout each semester. Each week, the course consisted of three 50-minute lectures and one 50-minute tutorial section where the University of Washington *Tutorials in Introductory Physics* [31] 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 pre-lecture 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 3,511 FMCE pretest responses (74% male) and 3,016 FMCE post-test responses (73% male).

Sample 2: Sample 2 was collected during 13 semesters 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” [216]. 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 [173]. The students in Sample 2, who were primarily engineering majors (85%), were enrolled in the introductory, calculus-based mechanics course. Only the students who completed the course for a grade and completed both the FMCE pretest and FMCE post-test were included. The Sample 2 dataset included 3,956 pretest responses (80% male) and 3,719 post-test responses (80% male).

The instructional environment for Sample 2 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 the fall 2011 and spring 2015 semesters, the LA program [36], that was presented in Chapter 4, was implemented as a tool to improve conceptual understanding of students in the introductory calculus-based sequence (refer to Sections 4.3 and 4.4).

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, the course was team-taught by a pair of experienced educators. The course consisted of three one-hour lectures and one three-hour laboratory each week. 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.

As such, the two samples are drawn from two diverse institutions with Sample 1 having the most academically prepared students measured by ACT scores and Sample 2 having the least well-prepared students.

FMCE Scoring

A modified scoring method for the FMCE proposed by Thornton *et al.* was employed in this study [212]. 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 [22, 217]. 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 the following analysis, the method of an 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).

		Male Students		Female Students	
	N	N	$(M \pm SD)\%$	N	$(M \pm SD)\%$
Sample 1					
FMCE Pretest	3511	2607	45 ± 28	904	30 ± 22
FMCE Post-test	3016	2192	74 ± 26	824	59 ± 28
Sample 2					
FMCE Pretest	3956	3146	25 ± 19	810	20 ± 14
FMCE Post-test	3719	2947	53 ± 28	772	41 ± 24

Table 6.5: FMCE pretest and post-test averages for Sample 1 and 2. Averages are reported as percentages.

Bonferroni Correction

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. For this analysis, 30 statistical tests were used for each analysis, adjusting $p < 0.05$ to $p < 0.0017$, $p < 0.01$ to $p < 0.0003$, and $p < 0.001$ to $p < 0.00003$.

6.2.4 Results

The differences in performance between male and female students were measured with t-tests. Cohen's d was used to characterize the effect size for each test. Similar to Section 6.1, item fairness will be examined graphically and with DIF analysis.

Table 6.5 presents the FMCE pretest and post-test averages for Sample 1 and Sample 2. In Sample 1, male students outperformed female students 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 2, significant gender differences were detected on both the FMCE pretest and post-test; however, these differences were smaller than those of Sample 1. Male students outperformed female students 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.

Difficulty and Discrimination

The problematic items in the FMCE for Sample 1 and Sample 2 are presented in Table 6.6. Table 6.7 presents the CTT item-level statistics for all of the FMCE items. For Sample 1, nearly half of the items on the FMCE pretest were problematic for female students with $P < 0.2$ except for items 40 and 43 with $P > 0.8$. Fewer problematic items were identified for male students; items 36 and 38 ($P < 0.2$) and items 40, 42, and 43 ($P > 0.8$) were problematic for men. For female students in Sample 1, only items 40 and 43 were problematic post-instruction ($P > 0.8$). While the number of problematic items decreased for female students from pretest to post-test, male students had more problematic items after instruction. All of the items identified as problematic for male students post-instruction had a difficulty of $P > 0.8$.

Gender	Pre/Post	Problematic Items
		Sample 1
Female	Pre	2, 4, 8_10, 11_13, 14, 17, 18, 19, 20, 27_29, 36, 38, 40, 43
	Post	40, 43
Male	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 2
Female	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
Male	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

Table 6.6: CTT problematic items with $P < 0.2$, $P > 0.8$, or $D < 0.2$ for the FMCE.

	Sample 1						Sample 2					
#	P_M	P_F	D_M	D_F	ϕ	$\Delta\alpha_{MH}$	P_M	P_F	D_M	D_F	ϕ	$\Delta\alpha_{MH}$
1	.72±.01	.55±.02	.74±.02	.81±.03	.15^c	.10	.41±.01	.27±.02	.88±.01	.59±.04	.12^c	.26
2	.70±.01	.53±.02	.70±.02	.86±.03	.15^c	-.04	.38±.01	.24±.02	.89±.01	.66±.04	.12^c	.24
3	.74±.01	.52±.02	.67±.02	.75±.03	.21^c	-1.01^b	.51±.01	.29±.02	.81±.01	.60±.04	.18^c	-1.31^c
4	.68±.01	.50±.02	.77±.02	.80±.03	.17^c	-.19	.40±.01	.25±.02	.84±.01	.59±.04	.13^c	-.10
7	.74±.01	.53±.02	.61±.02	.72±.04	.19^c	-.88^a	.53±.01	.32±.02	.76±.02	.63±.04	.17^c	-1.25^c
8_10	.63±.01	.38±.02	.83±.02	.87±.02	.23^c	-1.27^c	.25±.01	.10±.01	.74±.02	.33±.04	.14^c	-1.16
11_13	.75±.01	.52±.02	.70±.02	.86±.02	.22^c	-1.19^c	.37±.01	.18±.01	.83±.01	.51±.04	.16^c	-1.25^b
14	.69±.01	.54±.02	.84±.02	.87±.02	.13^c	.96^a	.40±.01	.26±.02	.94±.01	.75±.04	.12^c	.59
16	.71±.01	.56±.02	.86±.02	.86±.02	.14^c	.99	.43±.01	.28±.02	.92±.01	.72±.04	.12^c	.24
17	.59±.01	.45±.02	.85±.02	.82±.03	.13^c	.75	.34±.01	.21±.01	.84±.01	.67±.04	.11^c	.55
18	.65±.01	.47±.02	.88±.01	.89±.02	.17^c	.42	.37±.01	.23±.02	.91±.01	.64±.05	.12^c	.35
19	.65±.01	.44±.02	.88±.01	.84±.03	.20^c	-.39	.38±.01	.23±.02	.88±.01	.59±.05	.12^c	.14
20	.64±.01	.48±.02	.79±.02	.64±.04	.15^c	.07	.39±.01	.24±.02	.78±.02	.43±.05	.12^c	-.27
21	.72±.01	.50±.02	.70±.02	.82±.03	.21^c	-1.09^b	.41±.01	.25±.02	.82±.01	.62±.04	.14^c	-.49
22	.85±.01	.71±.02	.50±.03	.71±.03	.15^c	-.32	.70±.01	.57±.02	.76±.02	.84±.03	.12^c	-.24
23	.77±.01	.64±.02	.65±.02	.85±.02	.13^c	.56	.59±.01	.48±.02	.86±.01	.86±.02	.09^c	.44
24	.83±.01	.74±.02	.51±.02	.71±.04	.10^c	.72	.67±.01	.56±.02	.81±.02	.85±.03	.09^c	.36
25	.75±.01	.55±.02	.68±.02	.80±.03	.20^c	-.87^a	.56±.01	.41±.02	.84±.01	.77±.04	.13^c	-.26
26	.84±.01	.74±.02	.46±.03	.69±.03	.12^c	.28	.70±.01	.58±.02	.77±.02	.83±.03	.10^c	.08
27_29	.77±.01	.53±.02	.65±.02	.89±.02	.23^c	-1.50^c	.49±.01	.26±.02	.84±.01	.64±.03	.19^c	-1.66^c
30	.76±.01	.71±.02	.56±.02	.50±.04	.05	1.05^c	.64±.01	.63±.02	.58±.02	.50±.04	.01	1.09^c
31	.81±.01	.75±.01	.46±.02	.50±.04	.07^b	.70	.69±.01	.67±.02	.58±.02	.53±.04	.02	1.03^c
32	.78±.01	.70±.02	.55±.02	.63±.04	.08^c	.70^b	.65±.01	.61±.02	.69±.02	.68±.04	.03	1.10^c
34	.78±.01	.68±.02	.55±.02	.67±.03	.10^c	.41	.63±.01	.57±.02	.73±.02	.67±.04	.05	.86^a
36	.47±.01	.40±.02	.77±.02	.62±.04	.07^a	.89^b	.30±.01	.29±.02	.56±.02	.51±.04	.01	1.39^c
38	.47±.01	.38±.02	.79±.02	.63±.04	.08^c	.71	.31±.01	.28±.02	.56±.02	.52±.04	.03	1.03^b
40	.92±.01	.86±.01	.23±.02	.33±.04	.08^c	-.08	.94±.00	.86±.01	.16±.01	.32±.04	.12^c	-1.62^c
41	.84±.01	.73±.02	.37±.02	.49±.04	.12^c	-.30	.77±.01	.65±.02	.42±.02	.39±.04	.11^c	-.63
42	.90±.01	.80±.01	.28±.02	.42±.04	.13^c	-.68	.87±.01	.76±.02	.31±.02	.46±.04	.12^c	-1.03^b
43	.93±.01	.90±.01	.19±.02	.26±.03	.05	.48	.94±.00	.90±.01	.12±.01	.19±.03	.05^a	-.55

Table 6.7: CTT difficulty and discrimination and DIF $\Delta\alpha_{MH}$ for each FMCE item. The significant levels have been Bonferroni corrected: "a" denotes $p < 0.0016$, "b" denotes $p < 0.0003$, and "c" denotes $p < 0.00003$.

The results of the FMCE for Sample 2 were fairly similar to those in Sample 1. In Sample 2, however, both male and female students 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 male and female students in Sample 2. Items 40 and 43 continued to be problematic for students on the FMCE post-test. For female students, in addition to items 40 and 43, two of the three clustered items identified by Thornton *et al.* [212], 8_10 and 11_13, were problematic ($P < 0.2$). For male students, in addition to items 40 and 43, item 42 was a problematic item with $P > 0.8$.

Graphical Analysis

Similar to Section 6.1.4, Figure 6.5 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 one standard deviation in each direction.

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 ϕ coefficient was calculated for each item to characterize the effect size. Post-instruction, all items except for items 30 and 43 were significantly different for male and female students with female students scoring lower; however, none of the items showed more than a small effect size.

The FMCE pretest and post-test results for Sample 2 are presented in Fig. 6.5(c) and (d). The results were generally similar to the Sample 1 results. The Sample 2 pretest scores were substantially lower than the Sample 1 pretest scores which may have produced

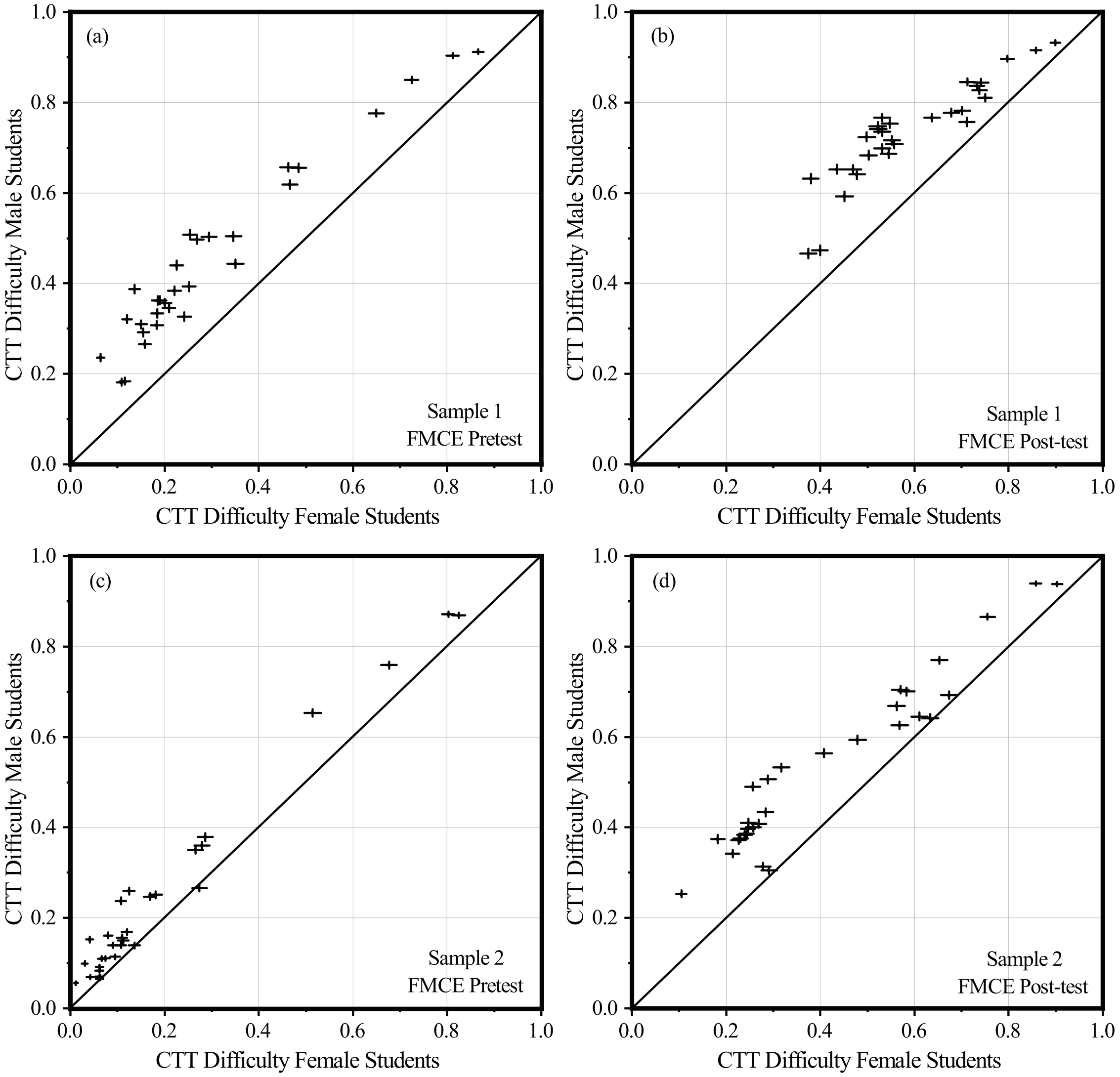

Figure 6.5: CTT difficulty results for the FMCE. The top two panels are Sample 1 (a) FMCE pretest and (b) FMCE post-test. The bottom two panels are Sample 2 (c) FMCE pretest and (d) FMCE post-test. A line of slope one is drawn to allow comparison of male and female difficulty. Error bars represent one standard deviation in each direction.

the clustering near the fairness line at scores less than 25% seen in Fig. 6.5(c). After instruction, the overall item difficulties for male and female students increased; however, most of the items were still significantly different for male and female students. Only items 30, 31, 32, 34, 36, and 38 were not significantly different for male and female students. 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. Unlike the previous work on the FCI, there was 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.

DIF Analysis

Table 6.8 presents the FMCE post-test $\Delta\alpha_{MH}$ 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 male students. With a value of $\Delta\alpha_{MH} = -1.50$, 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						
Item	P_M	P_F	D_M	D_F	ϕ	$\Delta\alpha_{MH}$
3	.74±.01	.52±.02	.67±.02	.75±.03	$.21^c$	-1.01^b
8_10	.63±.01	.38±.02	.83±.02	.87±.02	$.23^c$	-1.27^c
11_13	.75±.01	.52±.02	.70±.02	.86±.02	$.22^c$	-1.19^c
21	.72±.01	.50±.02	.70±.02	.82±.03	$.21^c$	-1.09^b
27_29	.77±.01	.53±.02	.65±.02	.89±.02	$.23^c$	-1.50^c
30	.76±.01	.71±.02	.56±.02	.50±.04	.05	1.05^c

Sample 2						
Item	P_M	P_F	D_M	D_F	ϕ	$\Delta\alpha_{MH}$
3	.51±.01	.29±.02	.81±.01	.60±.04	$.18^c$	-1.31^c
7	.53±.01	.32±.02	.76±.02	.63±.04	$.17^c$	-1.25^c
8_10	.25±.01	.10±.01	.74±.02	.33±.04	$.14^c$	-1.16
11_13	.37±.01	.18±.01	.83±.01	.51±.04	$.16^c$	-1.25^b
27_29	.49±.01	.26±.02	.84±.01	.64±.03	$.19^c$	-1.66^c
30	.64±.01	.63±.02	.58±.02	.50±.04	.01	1.09^c
31	.69±.01	.67±.02	.58±.02	.53±.04	.02	1.03^c
32	.65±.01	.61±.02	.69±.02	.68±.04	.03	1.10^c
36	.30±.01	.29±.02	.56±.02	.51±.04	.01	1.39^c
38	.31±.01	.28±.02	.56±.02	.52±.04	.03	1.03^b
40	.94±.00	.86±.01	.16±.01	.32±.04	$.12^c$	-1.62^c
42	.87±.01	.76±.02	.31±.02	.46±.04	$.12^c$	-1.03^b

Table 6.8: CTT difficulty and discrimination and DIF $\Delta\alpha_{MH}$ for the FMCE post-test items with small to moderate or large DIF. The significant levels have been Bonferroni corrected: "a" denotes $p < 0.0016$, "b" denotes $p < 0.0003$, and "c" denotes $p < 0.00003$.

Sample 1 presented in Table 6.8 (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 male students.

The results for Sample 2 were similar; however, in addition to item 27_29, item 40 also had large DIF with an advantage to male students. Items 3, 7, 8_10, 11_13, 30, 31, 32, 36, 38, and 42 demonstrated small to moderate DIF, half with an advantage to female students, half to male students.

Because large DIF items influence the overall test score, the identification of items with either large or small to moderate 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 these items reduced the gender gap in FMCE post-test scores by 2.5%. For Sample 2, 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 2 was reduced by 1.1%.

6.2.5 Discussion

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

RQ1: Are there items in the FMCE which CTT would identify as problematic? Are the problematic items the same for male and female students? 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 2. These findings supported those of Talbot [210] and Ishimoto [211] 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 [210] and Ishimoto [211], 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 [21]. In Section 6.1, FCI item 15 was identified as problematic pre-instruction.

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 male and female students in both Sample 1 and Sample 2. Although this result also agreed with the work presented by Talbot in which items 40 and 43 remained easy FMCE items after instruction [210], 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*) [213], which assesses the concept of gravity as a constant downward force, was difficult for female students prior to physics instruction for both Sample 1 and Sample 2. In Sample 2, two of the three items (8_10 and 11_13) within this cluster remained difficult post-instruction.

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 [21]. The previous section identified item 13 as problematic for female students prior to physics instruction but not after instruction. 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.

RQ2: Are there items in the FMCE which are substantially unfair to men or women? Although the incoming pretest scores were somewhat different for Sample 1 and Sample 2, the overall result that the majority of the FMCE items were more difficult for female students 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 the FMCE were quite different than those in the previous section for the FCI. 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 post-instruction in either Sample 1 or Sample 2.

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 2, 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 male and female students in Sample 2.

The other two clusters that were defined by Thornton *et al.* [212], items 8_10 and 11_13, demonstrated small to moderate DIF against female students 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 female students.

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

In general, the results for the FMCE presented in this study were quite different than the results of a the previous analysis of the FCI. With this, the FMCE is substantially more gender fair than the FCI at the item-level.

6.2.6 Implications

This work demonstrated that the FMCE has few items with large DIF while the previous section showed that the FCI contains multiple large-DIF items. As such, institutions making decisions on the assessment of instructional practices should consider using the FMCE for mechanics courses. The previous section 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 Sample 1 from the previous section than the FMCE in either Sample 1 or Sample 2 in this work.

6.2.7 Conclusions

The previous section summarized an analysis on the item-level fairness between male and female students on the FCI. This section extended that research to the FMCE. The majority of the items were significantly more difficult for female students both pre- and post-instruction; however, few items stood out as being substantially unfair. There was only one item across both of the samples that demonstrated large DIF.

6.3 Gender Fairness within the Conceptual Survey of Electricity & Magnetism*

6.3.1 Introduction

There have been very few studies that have focused on the individual items of the CSEM; no item level analysis of the CSEM has been reported differentiated by gender. Maloney *et al.* reported that the difficulty of the items on the CSEM were between 0.1 and 0.9 [23]. 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 did 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 [218]. 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) [219]. One out of five students had an alternate model of Newton's 3rd law prior to any electricity and magnetism instruction; however, post-instruction these students are likely to change their understanding to the correct model.

*This section is combined with the previous section for the second part of the two-part submitted publication in Physical Review Physics Education Research.

6.3.2 Research Questions

- RQ1: Are there items in the the CSEM which CTT would identify as problematic? Are the problematic items the same for male and female students?
- RQ2: Are there items in the CSEM which are substantially unfair to men or women?

6.3.3 Methods

Samples

Sample 1: Sample 1 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 23-29 (25th to 75th percentile range) [173]. The university had a classification of "Highest Research Activity" for the entire period studied [216]. The overall undergraduate demographics were 77% White, 8% Hispanic, 5% African American, 2% Asian with other groups each 3% or less [173].

The course was taught and overseen by one lead instructor over the time period studied. The course consisted of two 50-minute lectures and two two-hour 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 post instruction. The score on the CSEM was counted toward the students' course grade. The aggregated dataset ($n_{pre} = 2,108$, $n_{post} = 2,014$) 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%) with the majority of the students enrolled in engineering majors (80%). The instructional environment was the same as the sample described in Chapter 5.

Sample 2: Sample 2 was collected during the same 13 semesters at the same institution as Sample 2 in Section 6.2; a large eastern land-grant university serving approximately 30,000 students. The students in this sample however were enrolled in the introductory, calculus-based electricity and magnetism course. Only the students who completed the courses for a grade and completed both the pretest and post-test were included. The Sample 2 dataset included 3,185 pretest responses (83% male) and 2,657 post-test responses (81% male) from the CSEM. The instructional environment was the same as Sample 2 described in the previous section.

Bonferroni Correction

A Bonferroni correction was applied to adjust for the inflation of Type I error rate. For this analysis, 32 statistical tests were performed, adjusting $p < 0.05$ to $p < 0.0016$, $p < 0.01$ to $p < 0.0003$, and $p < 0.001$ to $p < 0.00003$.

6.3.4 Results

		Male Students		Female Students	
	N	N	$(M \pm SD)\%$	N	$(M \pm SD)\%$
Sample 1					
CSEM Pretest	2108	1618	29 ± 11	490	25 ± 8
CSEM Post-test	2014	1552	65 ± 16	462	59 ± 16
Sample 2					
CSEM Pretest	3185	2642	27 ± 11	543	24 ± 9
CSEM Post-test	2657	2155	46 ± 18	502	41 ± 17

Table 6.9: CSEM pretest and post-test averages for Sample 1 and 2. Averages are reported as percentages.

Sample 1 and Sample 2 were analyzed to explore gender differences on the CSEM. Overall averages are presented in Table 6.9. For Sample 1, 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 2 were similar with male students outperforming female students by 4% on the CSEM pretest [$t(895) = 8.30$, $p < 0.001$, $d = 0.35$] and by 5% on the CSEM post-test [$t(780) = 6.06$, $p < 0.001$, $d = 0.29$] also with small effect sizes.

Difficulty and Discrimination

Gender	Pre/Post	Problematic Items
		Sample 1
Female	Pre	5, 7, 10, 11, 14, 15, 16, 20, 22, 23, 24, 25, 26, 28, 29, 31
	Post	1, 12, 23, 26, 31, 32
Male	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 2
Female	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
Male	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

Table 6.10: CTT problematic items with $P < 0.2$, $P > 0.8$, or $D < 0.2$ for the CSEM.

Table 6.10 presents the problematic items for Sample 1 and Sample 2 with item difficulty and item discrimination outside the desired ranges. Table 6.11 presents the CTT item level statistics for all of the CSEM items.

In Sample 1, all of the problematic pretest items for female students had $P < 0.2$, while the majority of the problematic pretest items for male students had $P < 0.2$ except for items 21 and 27 which had $D < 0.2$. The results for the Sample 2 pretest were similar; the majority

	Sample 1						Sample 2					
#	P_M	P_F	D_M	D_F	ϕ	$\Delta\alpha_{MH}$	P_M	P_F	D_M	D_F	ϕ	$\Delta\alpha_{MH}$
1	.87±.01	.87±.02	.20±.02	.21±.04	.01	.37	.73±.01	.70±.02	.37±.03	.50±.05	.04	.25
2	.67±.01	.65±.02	.32±.03	.33±.05	.02	.34	.39±.01	.38±.02	.35±.03	.39±.06	.07	.35
3	.83±.01	.69±.02	.33±.02	.41±.05	$.14^c$	-1.23^b	.80±.01	.71±.02	.34±.02	.26±.05	$.11^c$	-.58
4	.66±.01	.55±.02	.65±.02	.47±.05	$.09^b$	-.20	.57±.01	.51±.02	.60±.02	.48±.05	.07	.04
5	.55±.01	.41±.02	.65±.03	.54±.05	$.11^c$	-.48	.47±.01	.38±.02	.59±.03	.55±.05	$.10^b$	-.35
6	.83±.01	.68±.02	.33±.03	.55±.05	$.15^c$	-1.29^b	.68±.01	.53±.02	.59±.02	.58±.05	$.13^c$	$-.94^a$
7	.72±.01	.63±.02	.57±.03	.49±.05	$.08^a$	-.02	.50±.01	.41±.02	.79±.02	.77±.04	$.10^c$	-.07
8	.73±.01	.64±.02	.44±.03	.51±.05	$.08^b$	-.28	.63±.01	.51±.02	.53±.02	.60±.05	$.11^c$	-.60
9	.79±.01	.71±.02	.39±.03	.47±.05	$.08^b$	-.33	.59±.01	.45±.02	.47±.03	.54±.05	$.12^c$	$-.93^a$
10	.56±.01	.45±.02	.60±.03	.46±.06	$.09^c$	-.31	.44±.01	.31±.02	.63±.02	.46±.05	$.11^c$	-.74
11	.34±.01	.33±.02	.31±.03	.45±.05	.01	.47	.35±.01	.31±.02	.51±.02	.54±.05	.06	.33
12	.90±.01	.85±.02	.17±.02	.24±.04	.07	-.62	.81±.01	.72±.02	.34±.02	.40±.05	$.09^a$	-.68
13	.78±.01	.72±.02	.37±.03	.44±.05	.06	-.01	.34±.01	.24±.02	.47±.03	.31±.05	$.09^a$	-.63
14	.31±.01	.26±.02	.36±.03	.37±.05	.05	-.05	.15±.01	.15±.02	.11±.02	.04±.05	.03	.28
15	.64±.01	.57±.02	.50±.03	.43±.06	.06	.02	.36±.01	.27±.02	.58±.02	.46±.05	$.08^a$	-.24
16	.62±.01	.54±.02	.51±.03	.49±.06	.07	-.12	.30±.01	.29±.02	.45±.02	.42±.06	.07	.63
17	.77±.01	.73±.02	.37±.03	.30±.05	.05	.01	.46±.01	.39±.02	.61±.02	.52±.05	.06	-.02
18	.68±.01	.72±.02	.24±.03	.34±.05	.04	.91	.55±.01	.57±.02	.23±.03	.36±.06	.07	.55
19	.85±.01	.78±.02	.32±.03	.34±.05	$.08^b$	-.50	.49±.01	.45±.02	.66±.02	.59±.05	.05	.33
20	.38±.01	.50±.02	.32±.03	.49±.06	$.11^c$	1.93^c	.19±.01	.23±.02	.22±.02	.31±.05	.04	.92
21	.74±.01	.74±.02	.39±.03	.32±.06	.00	.64	.25±.01	.25±.02	.23±.03	.10±.06	.08	.35
22	.41±.01	.37±.02	.36±.03	.32±.06	.03	.13	.38±.01	.39±.02	.10±.03	.08±.06	.02	.20
23	.86±.01	.84±.02	.27±.03	.37±.05	.03	.42	.61±.01	.55±.02	.63±.03	.57±.05	.05	.06
24	.50±.01	.39±.02	.51±.03	.44±.06	$.09^b$	-.43	.37±.01	.33±.02	.55±.02	.54±.05	.05	.33
25	.60±.01	.52±.02	.58±.03	.44±.06	.06	.03	.42±.01	.39±.02	.58±.02	.49±.05	$.10^c$	.45
26	.89±.01	.83±.02	.28±.02	.38±.05	$.08^b$	-.54	.68±.01	.58±.02	.63±.02	.66±.04	$.09^a$	-.28
27	.70±.01	.69±.02	.41±.03	.32±.06	.01	.59	.24±.01	.21±.02	.33±.03	.27±.05	.06	.17
28	.65±.01	.55±.02	.28±.03	.32±.06	$.08^a$	-.47	.63±.01	.56±.02	.36±.03	.44±.06	.07	-.25
29	.51±.01	.39±.02	.65±.03	.51±.06	$.10^c$	-.38	.21±.01	.14±.02	.35±.02	.23±.05	$.13^c$	-.65
30	.70±.01	.68±.02	.43±.03	.31±.06	.02	.50	.53±.01	.52±.02	.45±.03	.36±.06	.04	.43
31	.28±.01	.18±.02	.43±.03	.39±.05	$.09^b$	-.51	.13±.01	.09±.01	.26±.02	.15±.04	.07	-.31
32	.50±.01	.48±.02	.20±.03	.08±.06	.01	.08	.37±.01	.43±.02	.25±.03	.33±.06	$.08^a$	1.02^b

Table 6.11: CTT difficulty and discrimination and DIF $\Delta\alpha_{MH}$ for each CSEM item. The significant levels have been Bonferroni corrected: "a" denotes $p < 0.0016$, "b" denotes $p < 0.0003$, and "c" denotes $p < 0.00003$.

of problematic items had $P < 0.2$ for both male and female students, except for item 4 for female students and item 21 for male students which had $D < 0.2$. Overall, male and female students demonstrated little incoming knowledge of electricity and magnetism in both samples.

Table 6.10 also presents the problematic CSEM post-test items for Sample 1 and Sample 2. Post-instruction the number of problematic items was reduced for both male and female students in both samples. Although there was very little commonality in the CSEM post-test problematic items between Sample 1 and Sample 2, within each of the samples there were many common problematic items between male and female students.

In the Sample 1 post-test, items 1, 12, 23, and 26 were problematic for both male and female students ($P > 0.8$). In addition, for male students, items 3, 6, and 19 also had $P > 0.8$. For female students, item 31 had $P < 0.2$ and item 32 had $D < 0.2$. Most of the problematic CSEM post-test items in Sample 1 had $P > 0.8$. Only one item was identified as too difficult for female students (item 31) and one item failed to discriminate between female students who know the material and those that do not (item 32).

In the Sample 2 post-test, items 14, 22 and 31 were problematic for male and female students post-instruction. Items 14 and 31 had $P < 0.2$ and item 22 had $D < 0.2$. The other problematic items were less consistent for male and female students. For male students, item 12 ($P > 0.8$) and item 20 ($P < 0.2$) were problematic. Item 29 ($P < 0.2$) and item 21 ($D < 0.2$) were problematic for female students.

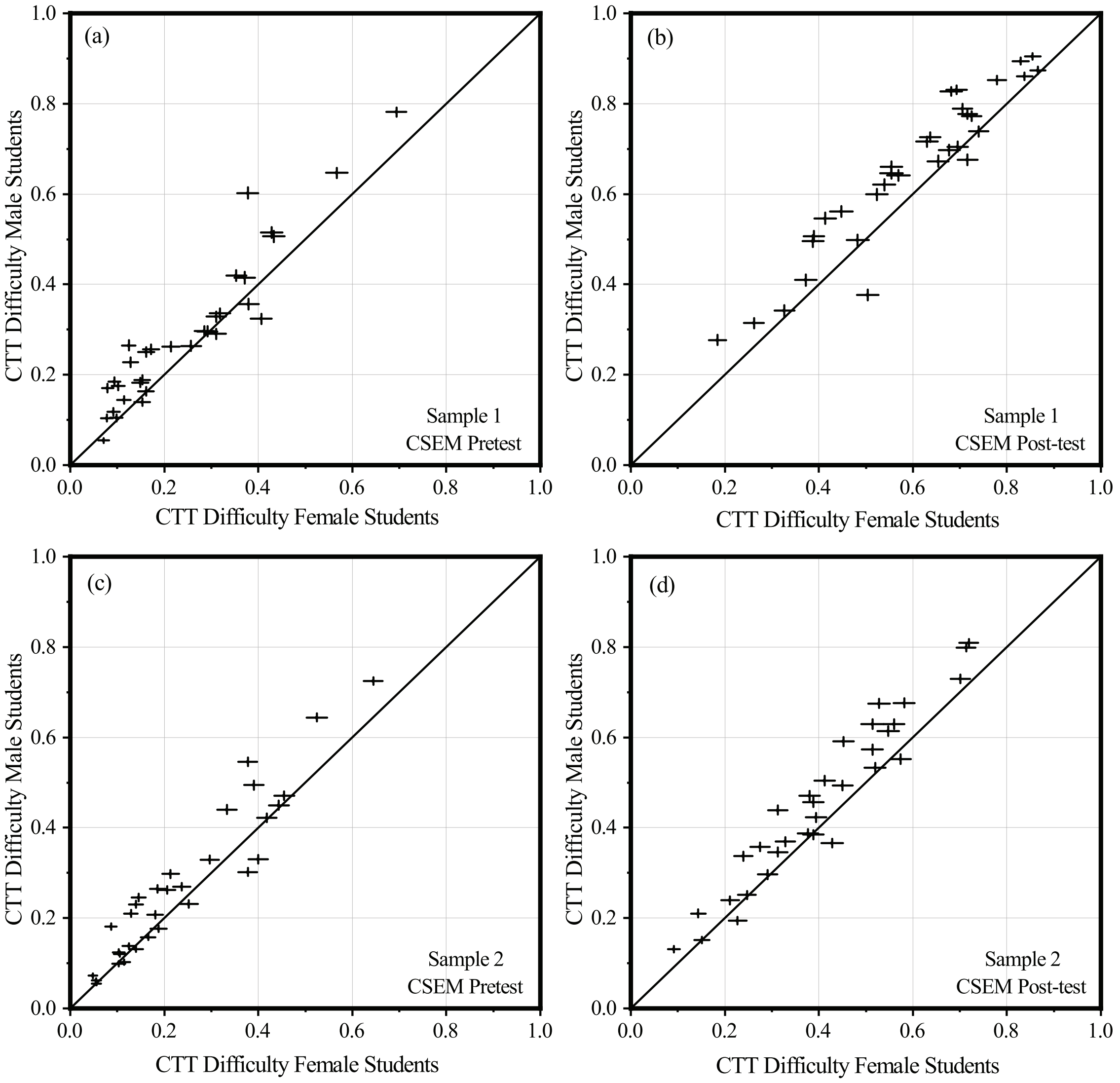

Figure 6.6: CTT difficulty results for the CSEM. The top two panels are Sample 1 (a) CSEM pretest and (b) CSEM post-test. The bottom two panels are Sample 2 (c) CSEM pretest and (d) CSEM post-test. A line of slope one is drawn to allow comparison of male and female difficulty. Error bars represent one standard deviation in each direction.

Graphical Analysis

Figure 6.6 plots the mean difficulties for the CSEM for men and women. Fig. 6.6(a) and 6.6(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 male nor female students could answer them.

In both CSEM post-test samples (Fig. 6.6(b) and (d)), the majority of the error bars do not overlap the fairness line; most items were significantly more challenging for female students. In Sample 1, there were two items that fell significantly below the fairness line and were more challenging to male students (items 18 and 20); however, in Sample 2, items more challenging for male students were closer to the fairness line. For Sample 1, a chi-squared test showed the difficulties for items 3, 5, 6, 20 and 29 were significantly different for male and female students with small effect sizes measured by the ϕ coefficient. For Sample 2, items 3, 5, 6, 7, 8, 9, 10, 25, and 29 were significantly different, also with small effect sizes.

DIF Analysis

Table 6.12 presents the items in the CSEM post-test that have either small to moderate or large DIF. In Sample 1, only item 20 demonstrated large DIF (unfair to male students), while two other items, 3 and 6, showed small to moderate DIF (unfair to female students). In Sample 2, only item 32 demonstrated small to moderate DIF; this item was moderately unfair to male students.

Sample 1						
Item	P_M	P_F	D_M	D_F	ϕ	$\Delta\alpha_{MH}$
3	$.74\pm.01$	$.52\pm.02$	$.67\pm.02$	$.75\pm.03$	$.14^c$	-1.01^b
6	$.83\pm.01$	$.68\pm.02$	$.33\pm.03$	$.55\pm.05$	$.15^c$	-1.29^b
20	$.38\pm.01$	$.50\pm.02$	$.32\pm.03$	$.49\pm.06$	$.11^c$	1.93^c
Sample 2						
Item	P_M	P_F	D_M	D_F	ϕ	$\Delta\alpha_{MH}$
32	$.37\pm.01$	$.43\pm.02$	$.25\pm.03$	$.33\pm.06$	$.08^a$	1.02^b

Table 6.12: CTT difficulty and discrimination and DIF $\Delta\alpha_{MH}$ for the CSEM items with small to moderate or large DIF. The significant levels have been Bonferroni corrected: "a" denotes $p < 0.0016$, "b" denotes $p < 0.0003$, and "c" denotes $p < 0.00003$.

To construct an unbiased instrument, for each sample, items were iteratively removed and DIF was recalculated. Because item 20 was substantially unfair to male students in Sample 1, removing items 3, 6, and 20 increased the original gender gap by 0.1%. Removing item 32 increased the gender gap in Sample 2 by 0.4%.

6.3.5 Discussion

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

RQ1: Are there items in the CSEM which CTT would identify as problematic? Are the problematic items the same for male and female students? Problematic items were identified within the CSEM. The results for the CSEM were less consistent between Sample 1 and Sample 2 than between the two samples for the FMCE. Within each sample, the identified problematic items in the CSEM were fairly consistent between male and female students. 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 male and female students.

For the CSEM post-test, different problematic items were identified in Sample 1 and Sample 2. In Sample 1, items 1, 12, 23, and 26 were easy problems for both male and female students post instruction; these items went from being too difficult to too easy for both genders. In Sample 2, items 14, 22 and 31 were identified as problematic for both male and female students 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 male students and item 31 had a difficulty of $P < 0.2$ for female students. 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 1 and Sample 2 (Table 6.9).

RQ2: Are there items in the CSEM which are substantially unfair to men or women? The item fairness of the CSEM was 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 1 and item 32 in Sample 2) unfair to male students. Overall, the majority of the CSEM items were not significantly different in difficulty for men and women.

DIF analysis was also performed for the CSEM. In both Sample 1 and Sample 2, only one item, item 20, demonstrated large DIF; this item was biased toward female students.

In general, the results for the CSEM presented in this study were also quite different than the results of the previous analysis of the FCI. Chapter 5 examined gender differences in the CSEM and in other conceptual problems and found a 5% CSEM post-test gap and a 3% gap in both qualitative lab quiz and qualitative test questions. As such, no more than 2% of the gender gap could be the result of instrumental bias which is consistent with small number of high DIF CSEM items identified in Chapter 6, Section 6.3. With this, the CSEM is substantially more gender fair than the FCI at the item-level.

6.3.6 Implications

Like the FMCE, this work demonstrated that the CSEM has few items with large DIF. As such, institutions making decisions on the assessment of instructional practices should consider using the CSEM for electricity and magnetism courses. Future research on the BEMA may also identify it as fair, but this research is still to be done.

6.3.7 Conclusions

The previous sections performed an analysis on the item-level fairness between male and female students on the FCI and the FMCE. This section extended that research to the CSEM. 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 male students.

Chapter 7

An Overall Synthesis of the Gender Gap

Overall, the conclusions that can be drawn from the previous analyses taken together are stronger than those that are drawn from the studies taken individually. This chapter will synthesize the results from Chapters 5 and 6.

7.1 Introduction

In the previous chapter, item-level analysis showed a number of unfair items in the FCI but a substantially fewer unfair items in the FMCE or the CSEM; some FMCE and CSEM items performed differently for men and women but most items performed consistently with the overall difference in post-test score. However, DIF analysis cannot eliminate the possibility of overall instrumental unfairness; DIF can only detect differential fairness between items. The possibility of a general bias in the instruments shared approximately equally by all items still exists.

Chapter 5 demonstrated that the low female pretest scores (probably caused by the large number of problematic pretest items identified in Table 6.10) shifted the female pretest score distribution sufficiently that it substantially overlapped the pure guessing score distribution and, therefore, compared to male pretest scores, the pretest scores of women were less predictive of their post-test scores. The pretest was less valid for women than for men.

This chapter will combine the conclusions from Chapters 5 and 6 to further understand the overall the gender gap in the conceptual inventories. The analysis will first explore the possibility of a general bias in the conceptual inventories by using the same binning technique as in Fig. 5.3. In addition, the analysis will analyze post-test gender gaps for populations with equal preparation by constructing a valid pretest score (i.e., eliminating the problematic pretest items on each of the conceptual inventories) and further suggest that the overall gender gap can then be partitioned into an instrumental gender gap and a gender gap that is attributed from sources other than instrumental effects.

7.2 Research Questions

- RQ1: Are the differences in overall performance between male and female students on physics conceptual inventories dependent on the student's pretest score?
- RQ2: For the FCI, the FMCE, and the CSEM, how much of the overall gender gap can be attributed to differences of equally prepared students?

7.3 Methods

7.3.1 Samples

This chapter will only report the results for each Sample 1 from Sections 6.1, 6.2, and 6.3, which will be labeled FCI, FMCE, and CSEM, respectively. The FCI sample was collected at the same institution as the CSEM sample. Many students matriculated from the mechanics course which produced the FCI sample to the electricity and magnetism class that produced the CSEM sample and, as such, the student population was similar. Further, the classes that produced these samples shared the same structure and pedagogy. While different instructors led each class, both instructors were educators recognized by their institution for their teaching.

The FCI sample had $4,187$ matched pretest/post-test pairs, the FMCE had $2,744$ matched pretest/post-test pairs, and the CSEM had $1,804$ matched pretest/post-test pairs. A limited number of female students had a raw pretest score above 15 on the FCI, above 14 on the FMCE pretest and above 12 on the CSEM and, therefore, the analysis is restricted to students with pretest score below these thresholds. Table 7.1 shows the female post-test

FCI							
Bin	0-5	6-7	8-9	10-11	12-13	14-15	> 15
% Women	56	37	28	26	19	13	8
FMCE							
Bin	0-4	5-6	7-8	9-10	11-12	13-14	> 14
% Women	45	38	32	29	25	25	15
CSEM							
Bin	0-6	7-8	9-10	11-12	> 12		
% Women	32	29	22	18	8		

Table 7.1: The percentage of women in each pretest score bin for each sample. $N = 1,047$ students had a FCI pretest score > 15. $N = 909$ students had a FMCE pretest score > 14. $N = 250$ students had a CSEM pretest score > 12.

score distribution binned by pretest score for the FCI, FMCE, and the CSEM.

7.3.2 Modified Conceptual Inventories

For the analysis, each of the conceptual inventories pretest and post-test scores will be modified. The modifications will be based upon the results from the identified problematic items on the FCI, the FMCE, and the CSEM pretest and unfair items on the post-tests in Chapter 6. To construct valid pretest scores for each instrument, items that were identified as problematic on the respective pretests for either male or female students will be eliminated. The reduced post-test scores will eliminate the small to moderate and large DIF items identified by the DIF analyses, leaving no item-level unfairness on the instrument. Table 7.2 summaries the included items on each of the valid pretests and the reduced post-tests.

7.3.3 The Instrumental Gender Gap

Removing gender unfair items from the instruments lowers the probability that the overall gender gap stems from an instrumental factor. In addition, removing problematic

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

Table 7.2: The items included on the valid pretests and reduced post-tests for the FCI, the FMCE, and the CSEM.

items for male and female students that were identified on the pretest produces a pretest score that is equally valid for men and women; it measures men and women with equal accuracy.

This chapter will compare fair (reduced) post-test scores produced by removing small to moderate and large DIF items from the post-test to valid pretest scores produced by eliminating problematic items from the pretest. The gender gap in the conceptual inventory post-test scores can then be partitioned into an instrumental gender gap resulting from performance differences of equally prepared students (measured by valid pretest scores) and a gap resulting from other factors,

$$G_{red} = G_{inst} + G_{other}, \tag{7.1}$$

where G_{red} is the overall gender gap for the reduced conceptual inventory, G_{inst} is the mean of the gender gaps for the valid pretest bins, and G_{other} is the gender gap not attributable to instrumental effects. G_{inst} could result from a general post-test gender bias. It could also result from differing responses to how the post-test or pretest was applied or incentivized. The failure to detect a gender gap in the "TestQuant" environment for the CSEM (Figure 5.3) suggests that G_{inst} does not result in a differing response to post-testing conditions for the FCI or the CSEM. In both cases, post-test performance was incentivized by making the score part of the student's grade.

7.4 Results

This section will explore the instrumental gender gap in each of the conceptual inventories separately by a series of steps:

1. In parallel with Chapter 5, overall instrumental fairness will be analyzed by binning students by pretest score. This reproduces the binning already presented.

2. The unfair items identified in the DIF analysis in Chapter 6 will be removed and the reduced post-test scores for male and female students will be analyzed by binned pretest score.

3. The problematic pretest items identified in Chapter 6 will be eliminated to form a

valid pretest score for all students. The reduced post-test scores will be analyzed with respect to these new bins. Male and female reduced post-test scores will be analyzed using a linear regression model and the instrumental gender gap will be calculated by averaging the binned gender gaps.

To analyze the gender gap in each of the pretest bins in these figures, t-tests were used and Cohen's d effect sizes were calculated. A Bonferroni correction was applied to correct for the number of statistical tests performed.

7.4.1 FCI

Figure 7.1(a) presents the FCI post-test scores for male and female students with respect to binned FCI pretest scores. Overall, none of the bins demonstrated a significant difference in post-test scores between male and female students. This result is flawed because the pretest score mis-measures female students' prior preparation; it is not valid.

To analyze the gender gap for the reduced FCI post-test, Figure 7.1(b) plots the male and female reduced FCI post-test scores vs. the binned FCI pretest scores. The elimination of the large number of unfair items identified in the FCI resulted in the female students outperforming the male students on the reduced FCI post-test. With a Bonferroni correction, only one of the pretest bins showed a significant difference in reduced post-test performance between male and female students, bin 8-9 [$t(390) = 2.86, p < 0.01, d = 0.23$], with a small effect size.

After eliminating the common problematic items identified in Section 6.1, the student's valid pretest score was made up of 19 items. Figure 7.1(c) presents the reduced FCI post-test percentages for male and female student plotted against the valid FCI pretest bin. For

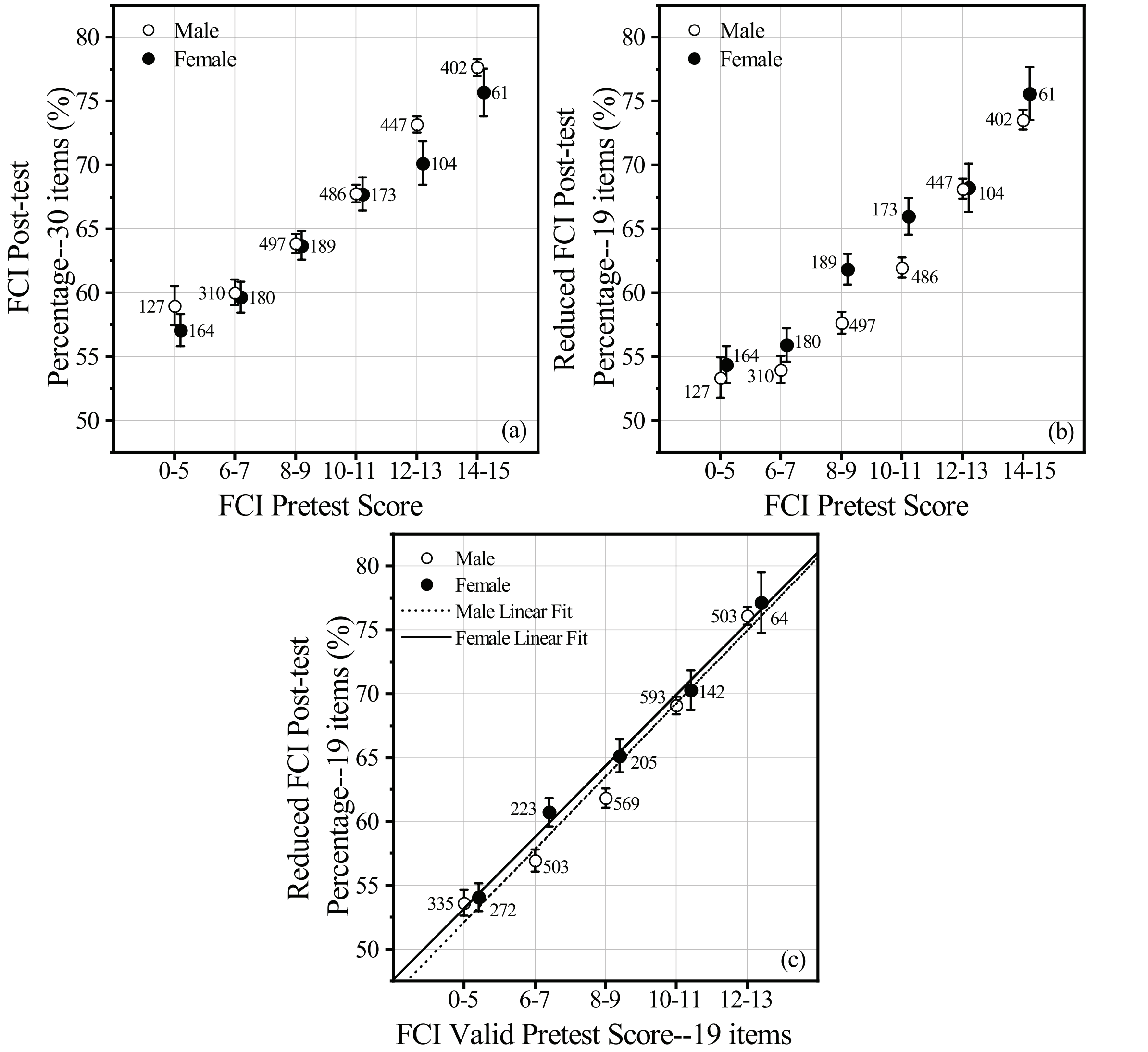

Figure 7.1: The FCI synthesis: (a) post-test percentage vs. pretest bin, (b) reduced post-test percentage vs. pretest bin, and (c) reduced post-test score vs. valid pretest bin. The number next to the data point is the number of students within each pretest range.

the five bins, only one bin, bin 6-7, showed a significant difference in reduced FCI post-test performance between male and female students with an advantage to female students $[t(478) = 2.68, p < 0.05, d = 0.21]$ with a small effect size. From Section 6.1, the overall gender gap for the reduced FCI was $G_{red} = 4.67\%$ and the instrumental gender gap was calculated to be $G_{inst} = -1.95\%$, leaving $G_{other} = 6.62\%$. This demonstrated that, after correcting for the mis-measurement of pretest score, female students outperform male students on the reduced FCI post-test.

The post-test scores were fitted for both male and female students separately with the female student fit indicated with the the solid line. The two fit lines are nearly parallel showing the instrumental bias is independent of preparation measured by pretest score, as expected. The vertical distance between the lines represents another measure of G_{inst}.

7.4.2 FMCE

Figure 7.2(a) plots the FMCE post-test percentage vs. the FMCE binned pretest score. With a Bonferroni correction, the difference in post-test performance between male and female students was not significant in any of the pretest bins.

In Section 6.2, DIF analysis identified a small number of items that were functioning differently for male and female students. Figure 7.2(b) presents the reduced FMCE, measured by a total of 23 items, vs. the binned FMCE pretest. Because of the small number of large DIF items detected in the FMCE, the post-test percentages of the FMCE for both male and female students were not as affected as they were for the reduced FCI. Correcting for Type I error rate, none of the gender gaps in any of the pretest bins were significant.

In Section 6.2, 15 items were identified as problematic prior to physics instruction for either male and female students; the valid FMCE pretest score was made up of the remaining 15 items of the FMCE. Figure 7.2(c) plots the reduced FMCE post-test percentage against valid FMCE pretest scores for male and female students. Although none of gender gaps in any of the valid pretest bins were significant, the linear fits demonstrated that, on average, male students outperform the female students. The instrumental gender gap was calculated to be $G_{inst} = 3.69\%$ and, with the reduced FMCE overall gender gap of $G_{red} = 11.76\%$

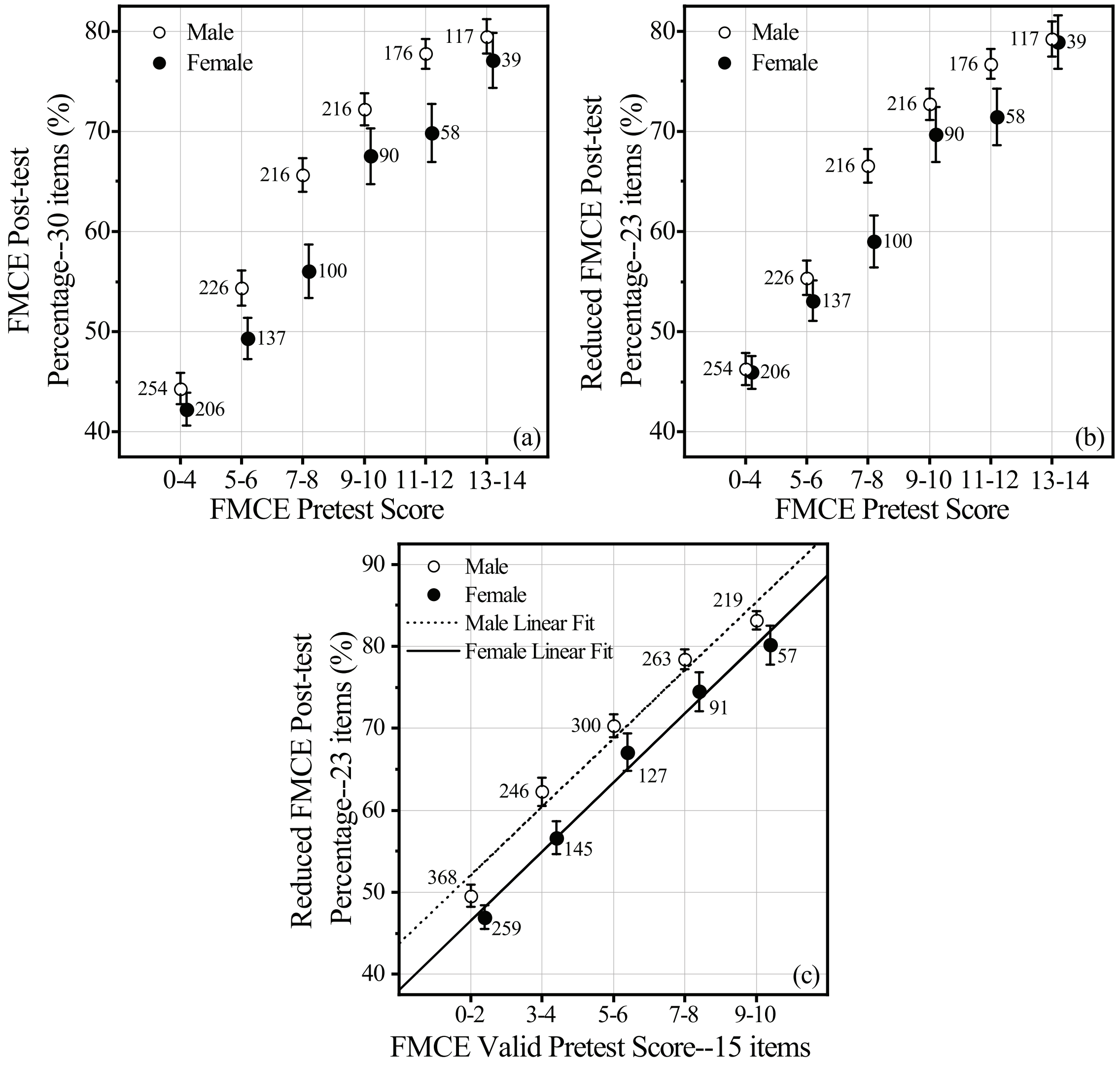

Figure 7.2: The FMCE synthesis: (a) post-test percentage vs. pretest bin, (b) reduced post-test percentage vs. pretest bin, and (c) reduced post-test score vs. valid pretest bin. The number next to the data point is the number of students within each pretest range.

reported in Section 6.2, the resulting gender gap due to other factors was $G_{other} = 8.07\%$. Approximately 70% ($= 100 \times G_{other}/G_{red}$) of the overall gender gap in the FMCE must be explained by differences other than instrumental bias leaving 30% of the gender gap explained by differences in performance of equally prepared students.

7.4.3 CSEM

Figure 7.3 presents a parallel analysis for the CSEM. With a Bonferroni correction, there was a significant difference in post-test performance between male and female students in bin 9-10 [$t(143) = 2.60$, $p < 0.05$, $d = 0.32$] with a small effect size.

In Section 6.3, to remove all item-level bias in the CSEM, only three items were eliminated to produce a reduced CSEM post-test consisting of 29 items. Figure 7.3(b) shows the gender gap of the reduced CSEM post-test by the students' CSEM pretest bin. Just as for the FMCE, since a small number of items were detected as highly unfair in the CSEM, the CSEM post-test scores for both male and female students were less affected than those of the FCI. For the reduced CSEM, differences in post-test scores were significant in bin 9-10 [$t(150) = 2.79$, $p < 0.05$, $d = 0.33$] and bin 11-12 [$t(65) = 2.78$, $p < 0.05$, $d = 0.47$].

Overall, Section 6.3 detected 18 problematic items on the CSEM pretest for either male or female students. Removing these items left 14 items to form a valid CSEM pretest score. Figure 7.3(c) plots the 29 items reduced CSEM post-test for each of the valid CSEM pretest bins. The linear fits show that male students outperformed female students. Only bin 6-7 showed a significant difference in post-test performance between male and female students [$t(240) = 4.27$, $p < 0.001$, $d = 0.42$]. The gender gap averaged over the valid pretest bins was $G_{inst} = 4.40\%$, leaving a non-instrumental gap of $G_{other} = 1.48\%$ ($G_{red} = 5.88\%$ from Section 6.3). For the CSEM, the majority of the overall gap (75%) resulted from differences in performance of equally prepared students.

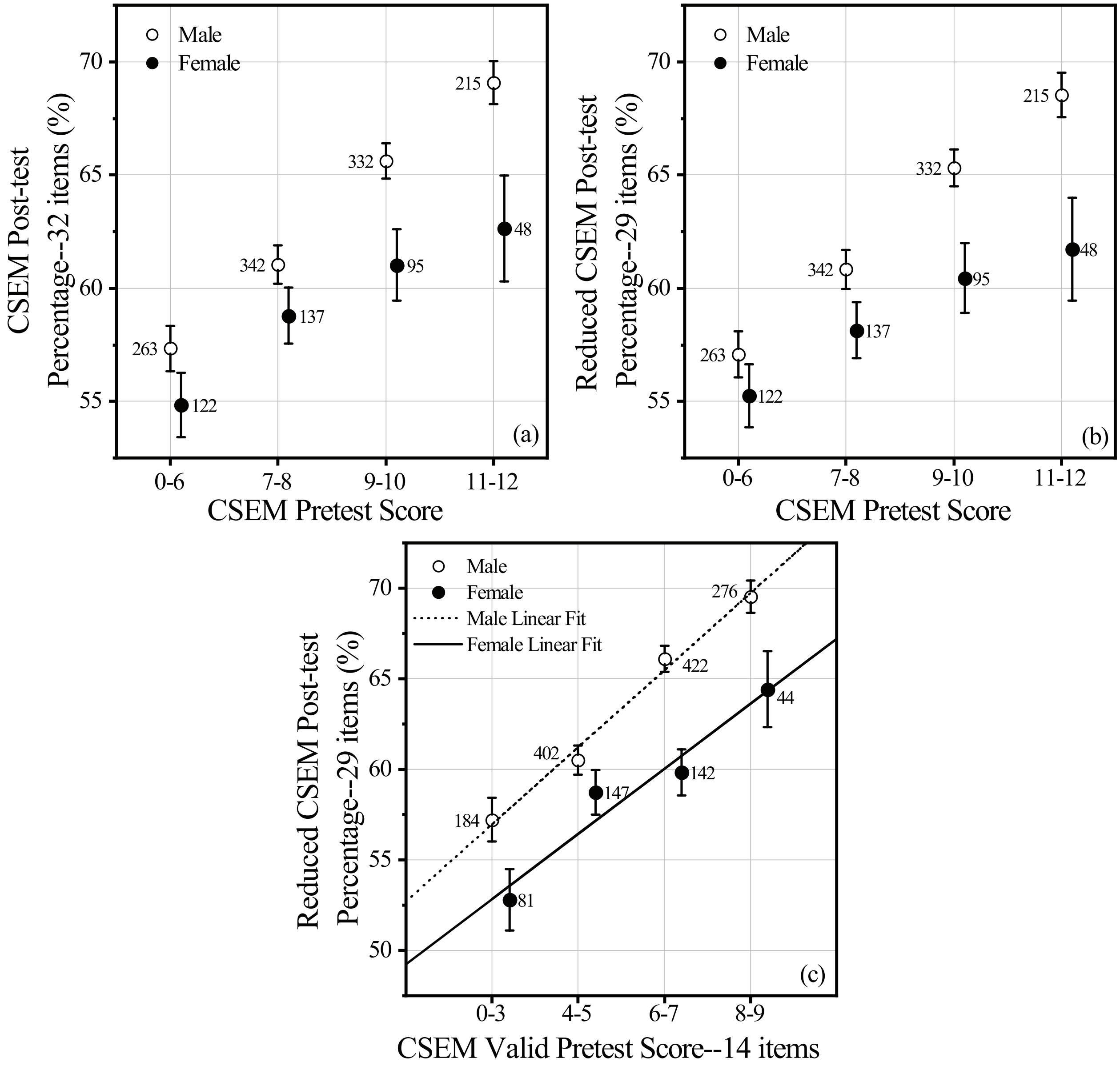

Figure 7.3: The CSEM synthesis: (a) post-test percentage vs. pretest bin, (b) reduced post-test percentage vs. pretest bin, and (c) reduced post-test score vs. valid pretest bin. The number next to the data point is the number of students within each pretest range.

7.5 Discussion and Conclusions

The above analysis showed that the 19-item FCI does not have an overall instrumental bias; however, even after correcting for differences in pretest scores, an instrumental bias remained on the reduced FMCE, and the reduced CSEM. A linear relation between valid pretest and reduced post-test was identified for men and women; in all samples, the linear relation for women was nearly parallel to the linear relation for men but displaced by G_{inst}. For the FCI, the instrumental gap was $G_{inst} = -1.95$, demonstrating that 100% of the overall gender gap in the reduced FCI can be explained by differences other than instrumental bias; the instrument is in fact slightly biased toward female students. However, this was not true for the FMCE and the CSEM. For equally prepared students, the FMCE and the CSEM showed a 4% instrumental gender gap. 30% and 75% of the overall gender gap in the FMCE and the CSEM, respectively, was explained by instrumental bias. Table 7.3 summarizes the results.

	G_{red} (%)	G_{inst} (%)	G_{other} (%)	Percentage due to Instrumental Bias	Percentage due to Other Factors
FCI	4.67	-1.95	6.62	0%	100%
FMCE	11.76	3.69	8.07	31%	69%
CSEM	5.88	4.40	1.48	75%	25%

Table 7.3: The reduced, instrumental, and non-instrumental gender gaps for the FCI, FMCE, and CSEM.

The FCI and CSEM samples were collected at the same institution. As such, differences between the FCI sample and the CSEM sample are unlikely to be the result of differences in either instruction or student population. The graphical plots, ϕ coefficients, and DIF analysis of the full FCI and CSEM (Figures 6.1 and 6.6(b)) were, however, dramatically different; a result of a substantial subset of large DIF items in the FCI and the small number identified in the CSEM. If large DIF items are removed to produce the 19-item instrument suggested by Section 6.1, the qualitative differences in Figure 6.1 and Figure 6.6(b) vanish. To determine the degree of similarity, one can count the number of ϕ coefficients that are smaller that 0.1, the threshold for small effect size. In the reduced 19-item FCI post-test, all items have $\phi < 0.1$. In the CSEM sample, removing the large DIF item, 3 of 31 items have $\phi \geq 0.1$. With the same student population and instructional environment, overall, the reduced FCI and the reduced CSEM behave similarly; however, the question remains as to the origin of the 4% instrumental bias seen on the CSEM. Because of the relation of the samples, it is unlikely to reside in differences in the students.

The partitioning of the overall gender gap provides an explanation of the differences in the ϕ coefficients between the FCI and the FMCE samples. While DIF analysis identified few unfair items in the FMCE compared to the FCI, the ϕ coefficients in the FMCE sample were much larger. However, qualitatively comparing the pretest results in Figures 6.4 and 6.5, the student population in the FMCE sample is more well prepared prior to physics instruction than the student population for the FCI sample; this is also seen in the pretest scores in Tables 6.1 and 6.5. Although the reduced post-test gap is much larger in the FMCE sample compared to the FCI sample (12% vs. 4%), 70% of the FMCE gap is attributable to factors other than the instrument. These "other" factors are almost certainly largely the result of the differences in preparation of male and female students shown in Table 7.1. Like the CSEM, the FMCE also shows a 4% instrumental gender gap. Further research should investigate the source of this gap.

The gender composition of the pretest bins in Table 7.1 clearly demonstrates substantial difference in preparation which may strongly influence the gap. While other factors may influence G_{other} beyond differential preparation, it seems likely that a large part of G_{other} results from prior preparation differences of male and female students; however, more research is needed to prove this hypothesis.

Overall, the percentage of the gender gap which could be attributed to instrumental effects varied widely by sample from 0% for the FCI to 4% for the FMCE and the CSEM. This analysis suggests that, for equally prepared student populations, the reduced FCI seems to be the most fair conceptual inventory to measure conceptual understanding of both male and female students. As such, institutions making decisions on the assessment of research-based instructional practices should consider using the reduced 19-item subset of the FCI.

Chapter 8

Racial and Ethnic Bias in the Force Concept Inventory[*]

The majority of this thesis work has explored the intersection of gender and physics. This chapter contains some additional work published within conference proceedings to examine the effect of race and ethnicity.

[*] This manuscript was published in the 2017 Physics Education Research Conference (PERC) proceedings. This work was supported in part by the National Science Foundation, PHY-0108787.

8.1 Introduction

While the gender gap in conceptual physics has been extensively studied, little research has explored whether these gaps extend to other underrepresented populations such as students of race or ethnicity different from the majority Caucasian population. In a recent commentary, Scherr lamented the unbalanced focus of PER on gender when many other populations are substantially underrepresented in physics [220]. A few studies have examined differences in post-test scores by race and ethnicity. Kost, Pollock and Finkelstein found that ethnicity was not a significant factor in predicting FMCE post-test scores; however, they warned that this could be due to the small number of students of non-majority ethnicity in their sample [72]. In another study, Kost-Smith, Pollock, and Finkelstein added ethnicity to their model when predicting Brief Electricity and Magnetism Assessment (BEMA) [24] post-test scores. They also found that ethnicity did not explain additional variability in post-test scores [74]. Hazari, Tai, and Sadler showed that there was a difference in introductory course performance, measured by grade, between Caucasian, African-American, and Hispanic students [70]. To the authors' knowledge, the validity and reliability of the FCI has not been explored for any racial or ethnic groups other than the majority Caucasian students.

The underrepresentation of women and minorities within Science, Technology, Engineering, and Mathematics (STEM) fields is a serious national concern. According the National Science Foundation, in 2012, 60% of bachelor's degrees awarded in science and engineering went to Caucasian students, 8.4% to African-American students, and 9.9% to Hispanic or Latino students [221]. Also in 2012, out of 5,557 awarded bachelor's degrees in physics, only 142 and 314 degrees in were given to African-American and Hispanic students respectively [221]. Although bachelor's degrees in STEM awarded to African-Americans grew from 1995 to 2004, an inverse relationship still exists between degree level and number awarded to African-Americans [222].

African-American and Hispanic men and women are not as likely to attend and complete college as Caucasian students. According to the *ACT Policy Report*, of the African-American and Hispanic students enrolled in a four-year college, 41% completed a degree within six years at the same institution compared to 59% of Caucasian students [223]. Braxton, Milem, and Sullivan demonstrated that race had a significant direct effect on intent to re-enroll in the upcoming semester [224]. Ishitani analyzed departure rates of transfer students; minority transfer students were 68% more likely to depart than their Caucasian transfer counterparts [225]. Toven-Lindsey *et al.* demonstrated that minority students who enrolled in an academic support program were more likely to persist in a science major [226]. Sociocultural factors that have been shown to be important for retaining women in STEM have also been investigated for underrepresented minorities [227, 228].

This study seeks to add to the sparse literature on the differences in conceptual physics performance between Caucasian, African-American, and Hispanic introductory physics students.

8.2 Research Questions

- RQ1: Are there differences in FCI post-test scores between male and female students?

- RQ2: Are there differences in course grades or FCI post-test scores between students of different race and ethnicity?

- RQ3: If a gender difference exists in FCI post-test scores, is this gender difference the same for students of all races and ethnicities?

8.3 Methods

This research was conducted in the first-semester, calculus-based mechanics course at a southern land-grant university serving approximately 25,000 students in the United States. The instructional environment was described within Sample 1 in Section 6.1.

From the fall 2006 semester to spring 2012 semester, 3,273 students completed the course for a grade. Of these students, 3,237 completed the FCI post-test. Students were also asked to self-report race and ethnicity as part of a survey administered at a different time than the FCI. Of the 3,237 students who completed the post-test, 2,038 reported their race or ethnicity. Race and ethnicity was collected as "Caucasian" ($n = 1,665$), "African-American" ($n = 85$), "Asian/Pacific Islander" ($n = 124$), "Hispanic" ($n = 82$), and "Other" ($n = 82$). Only students that reported African-American (4%), Caucasian (82%), or Hispanic (4%) were included in this study, leaving a sample size of 1,832 students (74% male students). For all analyses, gender was coded with female as 0 and male as 1 and Caucasian students were coded as the reference group. Course letter grades were measured on a four-point scale with A=4 and F=0. To explore the differences in FCI post-test scores between the various racial and ethnic groups ANOVA and hierarchical linear regression were employed.

8.4 Results

Table 8.1 summarizes FCI post-test scores and course letter grades; FCI scores are presented as a percentage. Significance levels were Bonferroni corrected to adjust for inflation of Type I error. Effect size was characterized by Cohen's d.

	Male Students		Female Students		
	N	$M \pm SD$	N	$M \pm SD$	d
African-American ($n = 85$)					
Physics Grade	61	2.9 ± 1.1	24	3.0 ± 0.7	0.11
FCI Post-test	61	63 ± 19	24	58 ± 13	0.32
Caucasian ($n = 1665$)					
Physics Grade	1241	3.4 ± 0.8	424	3.5 ± 0.7	0.17
FCI Post-test	1241	78 ± 15	424	70 ± 15	0.52
Hispanic ($n = 82$)					
Physics Grade	54	3.3 ± 0.8	28	3.1 ± 0.9	0.23
FCI Post-test	54	72 ± 15	28	66 ± 12	0.40

Table 8.1: Course letter grades and FCI post-test averages. Letter grades were measured on a four-point scale and FCI post-test averages are reported as percentages.

Differences between the students of different race and ethnicity were analyzed using a one-way between-subjects ANOVA. Results showed that there were significant differences in both letter grade $[F(2, 1829) = 16.67, p < 0.001]$ and FCI post-test score $[F(2, 1829) = 36.86, p < 0.001]$.

A posthoc analysis showed that there were significant differences in course grade between Caucasian and African-American students ($p < 0.01$) with a medium effect size of $d = 0.63$ and between Hispanic and African-American students ($p < 0.05$) with a small effect size of $d = 0.43$. There was no significant difference in course grade between Caucasian and Hispanic students. For the FCI post-test score, posthoc analysis showed significant differences between all races and ethnicities ($ps < 0.01$). There was a large effect size between

Caucasian and African-American students ($d = 0.90$), a medium effect size between Hispanic and African-American students ($d = 0.52$), and a small effect size between Caucasian and Hispanic students ($d = 0.36$).

Within each racial and ethnic group, differences by gender were analyzed using t-tests. For African-American students and Hispanic students there were no significant gender differences in either course grade or FCI post-test score. However, for Caucasian students, there was a significant gender difference in course grade [$t(814) = 3.13, p < 0.05, d = 0.17$] and FCI post-test score [$t(723) = 9.10, p < 0.001, d = 0.52$]. While significant gender differences were only measured for Caucasian students, the size of the differences were very similar between the three groups. An increase in sample size for African-American and Hispanic students may lead to the findings of significant gender differences.

To more thoroughly explore how the FCI post-test percentage was related to gender, race, and ethnicity, hierarchical linear regression was employed. Table 8.2 presents the results of this analysis.

In Model 1, course performance, measured by overall physics grade, was a significant predictor of FCI post-test average. This independent variable alone explained 27% of the variability in post-test scores.

Model 2 explored the relationship of gender and FCI post-test averages. There was a significant gender gap with male students outperforming female students by 7.76%; however, only 5% of the variance in FCI post-test average was explained by gender.

Model 3 examined the relationship of race and ethnicity with post-test average. This model compares African-American and Hispanic students to Caucasian students; Caucasian students form the baseline for the regression and the regression coefficient measures the

		Variables	B	SE	β	R^2_{adj}
Model 1		Physics Grade	10.18^c	0.39	0.52^c	$.27^c$
Model 2		Gender	7.76^c	0.83	0.49^c	$.05^c$
Model 3		African-Amer.	-14.03^c	1.73	-0.88^c	$.04^c$
		Hispanic	-5.64^b	1.76	-0.36^b	
Model 4		Physics Grade	9.84^c	0.39	0.50^c	$.28^c$
		African-Amer.	-9.07^c	1.51	-0.57^c	
		Hispanic	-4.48^b	1.52	-0.28^b	
Model 5	Step 1	Physics Grade	10.46^c	0.38	0.53^c	$.33^c$
		Gender	8.88^c	0.70	0.56^c	
	Step 2	Physics Grade	10.14^c	0.38	0.51^c	$.34^c$
		Gender	8.73^c	0.69	0.55^c	
		African-Amer.	-8.68^c	1.45	-0.55^c	
		Hispanic	-3.68^a	1.46	-0.23^a	
	Step 3	Physics Grade	10.16^c	0.38	0.52^c	.34
		Gender	9.13^c	0.73	0.58^c	
		African-Amer.	-6.88^c	2.71	-0.43^a	
		Hispanic	-0.09	2.52	-0.01	
		Afr.-Amer. × Gend.	-2.47	3.19	-0.16	
		Hisp. × Gend.	-5.40	3.09	-0.34	

Table 8.2: Hierarchical linear regression analysis predicting FCI post-test percentage. Female was coded as 0 and male as 1. B is the regression coefficient, SE the standard error, and β the regression coefficient normalizing the post-test percentage. Superscript "a" denotes $p < 0.05$, "b" denotes $p < 0.01$, and "c" denotes $p < 0.001$. The R^2_{adj} significance levels indicate the significance of the improved fit of the model over the model in which it is nested.

change with respect to this baseline. On average, African-American students (14.03%) and Hispanic students (5.64%) scored lower on the FCI post-test than Caucasian students. Race and ethnicity explained only 4% of the variability in FCI post-test score.

Model 4 explored the effect of race and ethnicity controlling for course performance. After controlling for physics grade, there were significant differences in FCI post-test scores between students of different race and ethnicity; African-American students (9.07%) and Hispanic students (4.48%) performed more weakly than Caucasian students. These differences were smaller than the uncorrected differences in Model 3. As such, some but not all of the differences in post-test scores for these students were explained by overall class performance.

Model 5 examined the relationship between gender, race, and ethnicity while controlling for course grade. Model 5–Step 1 identified a significant overall gender gap (8.88%) on the FCI post-test controlling for overall course performance measured by physics grade. Model 5–Step 1 significantly improved model fit ($p < 0.001$) over Model 1 and explained 33% of the variability in post-test scores. Model 5–Step 2 added race and ethnicity to the model. There was still a significant gender difference in post-test scores (8.73%) as well as significant differences between each racial and ethnic group. Controlling for course grade, African-American students still scored significantly lower than Caucasian students (8.68%) while Hispanic students also scored significantly lower than Caucasian students (3.68%). Although adding race and ethnicity to the model explained only an additional 1% of variability in post-test average, Model 5–Step 2 was a significantly better model than Model 5–Step 1 ($p < 0.001$).

Model 5–Step 3 introduced interactions between gender, race, and ethnicity. This model did not explain significantly more variability in post-test average and was not a significantly better model when compared to Model 5–Step 2. The interaction terms in this model were also not statistically significant. Although an overall gender gap exists in the FCI post-test score, this gender gap was the same for African-American, Caucasian, and Hispanic students correcting for course performance.

8.5 Discussion

RQ1: Are there differences in FCI post-test scores between male and female students? A gender gap of 7.76% was found in FCI post-test scores with men outperforming women (Table 8.2 Model 2). When controlling for course performance this difference in FCI post-test scores between male and female students was still found to be significant and relatively unchanged (8.88%) (Table 8.2 Model 5 – Step 1). Although this gender gap was lower than the overall average of 12% reported by Madsen, McKagen, and Sayre [63], the result was in the range of gender differences in FCI post-test scores in their review.

RQ2: Are there differences in course grades or FCI post-test scores between students of different race and ethnicity? The differences in course grades between Caucasian, African-American, and Hispanic students were similar to those presented previously [70]. Differences were also identified in FCI post-test scores. Controlling for course performance, these differences persisted but narrowed. Controlling for course grade, a 9.07% difference was measured between African-American students and Caucasian students and a difference of 4.48% between Hispanic students and Caucasian students (Table 8.2 Model 4). This study detected not only a gender gap in the FCI but also racial and ethnic differences not previously reported [72, 74]. Only some of the racial and ethnic gaps were explained by course grades while none of the gender gap was explained by grades.

RQ3: If a gender difference exists in FCI post-test scores, is this gender difference the same for students of all races and ethnicities? Comparison of Model 4 and Model 5–Step 2 in Table 8.2 showed that there was both an overall effect of gender (8.73%) and of race/ethnicity [African-American 8.68%; Hispanic 3.68%] which coexisted. The failure to find interactions between gender and race/ethnicity in Model 5–Step 3 suggests the gender gap is consistent across students of all races and ethnicities. The gender gap was neither localized in the majority Caucasian population, nor more or less severe for African-American or Hispanic students.

8.6 Implications and Limitations

The effect of gender was fairly orthogonal to the effect of race/ethnicity leading to the concern that interventions or modifications to the FCI instrument intended to narrow the gender gap might not help underrepresented students of non-majority race or ethnicity.

This work was performed at one institution and the results may be dependent on the student population or the instructional environment. The work relied on self-reported race and ethnicity information and forced students to identify a single race or ethnicity which may miscount multi-race students. In the future, research will focus on the reasons for the differences between the various racial and ethnic groups.

8.7 Conclusion

The gender gap on popular physics conceptual assessments has been thoroughly explored [63]. Race and ethnicity has been less studied. In this work, a gender gap (7.76%) was found on the FCI post-test. Differences between students by race and ethnicity were analyzed; Caucasian students outperformed both African-American and Hispanic students by 14.03% and 5.64%, respectively. Controlling for course performance measured by overall physics grade, the difference between male and female students was slightly increased. Differences between students by race and ethnicity were also significant after controlling for overall physics grade, but somewhat reduced. Although main effects of gender and race/ethnicity were present in this analysis, no significant race/ethnicity-by-gender interaction was measured. The gender gap was shared equally by students of all races and ethnicities.

Chapter 9

Future Work

Although this work suggested some of the overall gender gap in the conceptual inventories was the result of instrumental bias, additional research is required to understand the remaining performance differences to ensure that all students learn physics to the best of their ability. Future research plans revolve around the development of instructional pedagogies and assessment tools that can be used for improving learning for all students while at the same time proving additional support for underrepresented students. Some potential future and ongoing projects are outlined below:

- Continue to explore the gender gap by using additional variables (collected at WVU) to understand the latent CPP/NonQnt variable and thus understand the cause of male/female performance differences.
- Perform a similar fairness analysis on the Brief Electricity and Magnetism Assessment.
- Extend the fairness analysis to investigate the different experiences in physics of all underrepresented students (i.e., First Generation Status, Rural/non-rural, etc.).
- Qualitatively explore the 5 items that were identified as substantially unfair to female students in the FCI.
- Create adaptive instructional tools with an eye to inclusion.
- Construct and validate other assessment tools that target reformed pedagogies and materials making them sufficiently fine-grained that they allow instructors to identify specific areas of instruction that could be improved.

Chapter 10

Conclusion

With the introduction of research-based instruments such FCI [21], the FMCE [22], and the CSEM [23], differences in performance by gender began to be reported. The "gender gap" has been extensively studied in the mechanics inventories such as the FCI and the FMCE; however, much less work exists exploring the gender gap in the CSEM. On average, male students outperform female students by 13% on pretest scores and by 12% post instruction on the mechanics conceptual inventories and by 3.7% on pretest scores and by 8.5% on post-test scores on the electricity and magnetism conceptual inventories [63]. Many factors have been proposed that may influence the gender gap, from differences in background and preparation [72, 74] to various psychological and sociocultural effects, such as science anxiety [99] and stereotype threat [110]. A parallel but largely disconnected set of research has identified gender biased questions within the FCI [190, 201, 202]. This research has produced sporadic results and has only been performed on the FCI. Although the research into exploring the gender differences in physics is robust, researchers have yet to come to an agreement as to why a "gender gap" exists in the various conceptual inventories that are widely used in PER and/or how to reduce the gaps.

The motivation behind this work began first with seeking to understand the effects of the implementation of a LA program. Over the course of nine semesters, a strong oscillation of normalized gain between spring and fall semesters was measured. Hierarchical Linear Regression, controlling for spring/fall semester and standardized test scores (ACT/SAT), showed that the LA program increased normalized gain on the FMCE by approximately 20%; however, the effect of the 14 different instructors was much larger than the effect of the LA program. Throughout this analysis, male students outperformed women by 10%; an effect that was virtually unchanged by the addition of any other variable. Overall, this program had many departmental impacts and many valuable lessons were learned throughout the evaluation: (1) the program initially demonstrated a negative effect on learning because of the spring/fall fluctuation of student performance and preparation (ACT/SAT scores) which suggests that multiple semesters of control data are needed to understand educational reform, (2) different instructors produced different outcomes; therefore, it is critical to have coherent support and buy-in across all elements of the course, (3) the *Tutorials* and Learning Assistant program were introduced prior to the effort to produce coordinated instruction which suggests that introducing research-based materials before establishing consistent course structure may limit the effectiveness of reformed instruction, (4) the PER instruments that were used to monitor the courses did not allow the adopters to identify places where the innovations could be improved.

Knowing that little research exploring the overall gender gap in the electricity and magnetism conceptual inventories existed, Chapter 5 analyzed differences in performance between men and women on the CSEM. The data for this study was unique; it collected item-level data for all assignments in a physics course for 5 years. Each item was coded as either a qualitative or quantitative physics problem. This provided additional sources of qualitative questions beyond the CSEM and a control sample of quantitative questions. An overall gender gap existed in all of the qualitative environments: CSEM post-test, qualitative lab quiz questions, and qualitative test questions. However, male and female students performed equally on quantitative test questions. The failure to find performance differences on quantitative test questions suggested that the gender gap cannot be explained by psychological mechanisms such as science anxiety or stereotype threat.

The gender gaps in the qualitative environments were dependent on CSEM pretest score. This difference between qualitative and quantitative performance suggested that an underlying variable could be used to explain the differences. SEM was used to extract a latent variable which was called Conceptual Physics Performance/Non-Quantitative (CPP/NonQnt). This variable was the part of the student's conceptual performance that was not explained by their quantitative test performance in the course. The correlation between CPP/NonQnt and CSEM pretest score was larger for male students than it was for female students, suggesting that the CSEM pretest is less predictive of CPP/NonQnt for women than for men. To investigate this further, a sequence of models was fit to the CSEM pretest distributions for both male and female students. For male students, the differences in fit suggested that the CSEM pretest could discriminate between male students who were guessing and those that had some prior knowledge of electricity and magnetism. However, for female students this wasn't the case and therefore, the pretest scores provided less information about female students.

The analysis in Chapter 5 also suggested that the CSEM was not intrinsically biased. This conclusion was later refined in Chapter 7. To further explore this conclusion, exploration of intrinsic bias in the conceptual instruments themselves was performed. In Chapter 6, Classical Test Theory and Item Response Theory were used to compare item difficulties between male and female students on the FCI, FMCE, and the CSEM.

For the FCI post-test scores, most of the items were unfair and there were 5 items that stood out as substantially unfair to female students. Differential Item Functioning analysis using three academically diverse populations detected 10 items with high gender bias; eliminating these 10 items reduced the overall gender gap by 50%. This analysis suggested an extension to the validity framework published by Jorion *et al.* [142] which would add an item fairness analysis using DIF analysis to the validation process.

The item fairness analysis was extended to the FMCE and the CSEM. The results for the FMCE and the CSEM were quite different than those from the FCI. For the FMCE and CSEM post-test scores, most of the items were significantly more challenging for female students than for male students; however, unlike in the FCI, no items stood out as substantially unfair in the FMCE or the CSEM.

To synthesize the results in Chapter 5 and 6, valid pretest scores and reduced post-test scores were constructed for each of the conceptual inventories. This analysis demonstrated that the amount of instrumental bias attributing to the overall reduced post-test gaps varied between the FCI, the FMCE, and the CSEM. The differences in reduced FCI post-test performance between male and female students can be fully explained by factors other than instrumental bias; however, only 70% and 30% of the overall gender gap is due to other factors on the FMCE and CSEM, respectively. Due to the demonstrated differences in preparation seen in Chapter 7, the "other" factors attributing to the gender gap are almost certainly a result of these differences; however, further research is needed to prove this hypothesis. In general, for equally prepared students, the reduced FCI seems to be the most fair instrument.

Overall, this work showed features of the PER conceptual instruments that were not equally fair for men and women. It demonstrated ways to remove this unfairness and suggested that institutions making decisions on the assessment of research-based instructional practices should consider these features. In the big picture, to bring more women and other underrepresented students into the STEM fields, science classrooms should ensure equal opportunity to show individual learning capabilities; without valid and fair assessment tools and effective pedagogies that acknowledge diversity, there is not a guarantee to equal educational access.

Bibliography

[1] J.L. Docktor and J.P. Mestre. Synthesis of discipline-based education research in physics. *Phys. Rev. Phys. Educ. Res.*, 10(2):020119, 2014.

[2] J.H. Larkin and G.C. Brackett. Mathematics pre-requisites: A mastery approach. *Am. J. Phys.*, 42(12):1089–1091, 1974.

[3] H.T. Hudson and W.R. McIntire. Correlation between mathematical skills and success in physics. *Am. J. Phys.*, 45(5):470–471, 1977.

[4] D. Liberman and H.T. Hudson. Correlation between logical abilities and success in physics. *Am. J. Phys.*, 47(9):784–786, 1979.

[5] A.B. Champagne, L.E. Klopfer, and J.H. Anderson. Factors influencing the learning of classical mechanics. *Am. J. Phys.*, 48(12):1074–1079, 1980.

[6] H.D. Cohen, D.F. Hillman, and R.M. Agne. Cognitive level and college physics achievement. *Am. J. Phys.*, 46(10):1026–1029, 1978.

[7] B. Inhelder and J. Piaget. *The Growth of Logical Thinking from Childhood to Adolescence: An Essay on the Construction of Formal Operational.* Basic, New York, NY, 1958.

[8] J. Nussbaum and J.D. Novak. An assessment of children's concepts of the earth utilizing structured interviews. *Sci. Educ.*, 60(4):535–550, 1976.

[9] L. Leboutet-Barrell. Concepts of mechanics among young people. *Phys. Educ.*, 11(7):462, 1976.

[10] Y. Waern. On the relationship between knowledge of the world and comprehension of texts assimilation and accommodation effects related to belief structure. *Scand. J. Psychol.*, 18(1):130–139, 1977.

[11] D.E. Trowbridge and L.C. McDermott. Investigation of student understanding of the concept of velocity in one dimension. *Am. J. Phys.*, 48(12):1020–1028, 1980.

[12] D.E. Trowbridge and L.C. McDermott. Investigation of student understanding of the concept of acceleration in one dimension. *Am. J. Phys.*, 49(3):242–253, 1981.

[13] P.C. Peters. Even honors students have conceptual difficulties with physics. *Am. J. Phys.*, 50(6):501–508, 1982.

[14] M. McCloskey, A. Caramazza, and B. Green. Curvilinear motion in the absence of external forces: Naive beliefs about the motion of objects. *Science*, 210(4474):1139–1141, 1980.

[15] A. Caramazza, M. McCloskey, and B. Green. Naive beliefs in sophisticated subjects: Misconceptions about trajectories of objects. *Cognition*, 9(2):117–123, 1981.

[16] R.F. Gunstone and R.T. White. Understanding of gravity. *Sci. Educ.*, 65(3):291–299, 1981.

[17] J. Clement. Students preconceptions in introductory mechanics. *Am. J. Phys.*, 50(1):66–71, 1982.

[18] L.C. McDermott. Research on conceptual understanding in mechanics. *Phys. Today*, 37:24–32, 1984.

[19] I.A. Halloun and D. Hestenes. The initial knowledge state of college physics students. *Am. J. Phys.*, 53(11):1043–1055, 1985.

[20] I.A. Halloun and D. Hestenes. Common sense concepts about motion. *Am. J. Phys.*, 53(11):1056–1065, 1985.

[21] D. Hestenes, M. Wells, and G. Swackhamer. Force Concept Inventory. *Phys. Teach.*, 30:141–158, 1992.

[22] 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(4):338–352, 1998.

[23] 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(S1):S12–S23, 2001.

[24] L. Ding, R. Chabay, B. Sherwood, and R. Beichner. Evaluating an electricity and magnetism assessment tool: Brief Electricity and Magnetism Assessment. *Phys. Rev. Phys. Educ. Res.*, 2(1):010105, 2006.

[25] E. Mazur. *Peer Instruction: A User's Manual.* Prentice Hall, Upper Saddle River, NJ, 1997.

[26] Physport. `https://www.physport.org`. Accessed 8/8/2017.

[27] C. Hieggelke and T. O'Kuma. The impact of physics education research on the teaching of scientists and engineers at two-year colleges. In *AIP Conference Proceedings*, volume 399, pages 267–288. AIP, 1997.

[28] L.C. McDermott and P.S. Shaffer. Research as a guide for curriculum development: An example from introductory electricity. Part I: Investigation of student understanding. *Am. J. Phys.*, 60(11):994–1003, 1992.

[29] P.S. Shaffer and L.C. McDermott. Research as a guide for curriculum development: An example from introductory electricity. Part II: Design of instructional strategies. *Am. J. Phys.*, 60(11):1003–1013, 1992.

[30] 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(1):64–74, 1998.

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

[32] L.C. McDermott. Oersted Medal Lecture 2001: "Physics Education Research-the key to student learning". *Am. J. Phys.*, 69(11):1127–1137, 2001.

[33] A. Elby, R.E. Scherr, T. McCaskey, R. Hodges, E.F. Redish, D. Hammer, and T. Bing. Open Source Tutorials in Physics sensemaking: Suite I. `http://umdperg.pbworks.com/w/page/10511239/Tutorials%20in%20Physics%20Sense-Making`. Accessed 1/23/2018.

[34] Michael C Wittmann, Richard N Steinberg, and Edward F Redish. Activity-based tutorials: Introductory physics, the physics suite, volume 1. *Activity-Based Tutorials: Introductory Physics, The Physics Suite, Volume 1, by Michael C. Wittmann, Richard N. Steinberg, Edward F. Redish, pp. 126. ISBN 0-471-48776-7. Wiley-VCH, April 2004.*, page 126, 2004.

[35] M.J. Wittmann, R.N. Steinberg, and E.F. Redish. *Activity-based Tutorials. Volume 2: Modern Physics.* John Wiley & Sons, 2005.

[36] 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(11):1218–1224, 2010.

[37] Learning Assistant Program. `http://laprogram.colorado.edu/`. Accessed 2/7/2018.

[38] P.M. Miller, J.S. Carver, A. Shinde, B. Ratcliff, and A.N. Murphy. Initial replication results of learning assistants in university physics. In *AIP Conference Proceedings*, volume 1513, pages 30–33. AIP, 2013.

[39] E.W. Close, J. Conn, and H.G. Close. Becoming physics people: Development of integrated physics identity through the Learning Assistant experience. *Phys. Rev. Phys. Educ. Res.*, 12(1):010109, 2016.

[40] E.F. Redish. *Teaching Physics with the Physics Suite.* Wiley, 2003.

[41] D.R. Sokoloff and R.K. Thornton. Using interactive lecture demonstrations to create an active learning environment. In *AIP Conference Proceedings*, volume 399, pages 1061–1074. AIP, 1997.

[42] M.D. Sharma, I.D. Johnston, H. Johnston, K. Varvell, G. Robertson, A. Hopkins, C. Stewart, I. Cooper, and R. Thornton. Use of interactive lecture demonstrations: A ten year study. *Phys. Rev. Phys. Educ. Res.*, 6(2):020119, 2010.

[43] L. Deslauriers, E. Schelew, and C. Wieman. Improved learning in a large-enrollment physics class. *Science*, 332(6031):862–864, 2011.

[44] J.E. Caldwell. Clickers in the large classroom: Current research and best-practice tips. *CBE-Life Sci. Educ.*, 6(1):9–20, 2007.

[45] C. Wieman and K. Perkins. Transforming physics education. *Phys. Today*, 58(11):36, 2005.

[46] F.M. Goldberg, S. Robinson, and V.K. Otero. *Physics & Everyday Thinking.* It's About Time, Herff Jones Educational Division, 2008.

[47] PET:Physics and everyday thinking. `http://cpucips.sdsu.edu/web/pet/`. Accessed 2/7/2018.

[48] F. Goldberg, V. Otero, and S. Robinson. Design principles for effective physics instruction: A case from physics and everyday thinking. *Am. J. Phys.*, 78(12):1265–1277, 2010.

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

[50] C. Henderson and M.H. Dancy. Barriers to the use of research-based instructional strategies: The influence of both individual and situational characteristics. *Phys. Rev. Phys. Educ. Res.*, 3(2):020102, 2007.

[51] S.J. Pollock and N.D. Finkelstein. Sustaining educational reforms in introductory physics. *Phys. Rev. Phys. Educ. Res.*, 4(1):010110, 2008.

[52] C. Singh. When physical intuition fails. *Am. J. Phys.*, 70(11):1103–1109, 2002.

[53] F. Reif and J.I. Heller. Knowledge structure and problem solving in physics. *Educ. Psychol.*, 17(2):102–127, 1982.

[54] E. Yerushalmi, C. Henderson, K. Heller, P. Heller, and V. Kuo. Physics faculty beliefs and values about the teaching and learning of problem solving. I. Mapping the common core. *Phys. Rev. Phys. Educ. Res.*, 3(2):020109, 2007.

[55] C. Henderson, E. Yerushalmi, V.H. Kuo, K. Heller, and P. Heller. Physics faculty beliefs and values about the teaching and learning of problem solving. II. Procedures for measurement and analysis. *Phys. Rev. Phys. Educ. Res.*, 3(2):020110, 2007.

[56] D. Huffman. Effect of explicit problem solving instruction on high school students' problem-solving performance and conceptual understanding of physics. *J. Res. Sci. Teach.*, 34(6):551–570, 1997.

[57] E. Brewe. Modeling theory applied: Modeling instruction in introductory physics. *Am. J. Phys.*, 76(12):1155–1160, 2008.

[58] C. Singh. Student understanding of quantum mechanics. *Am. J. Phys.*, 69(8):885–895, 2001.

[59] S.V. Chasteen and S.J. Pollock. A research-based approach to assessing student learning issues in upper-division electricity & magnetism. In *AIP Conference Proceedings*, volume 1179, pages 7–10. AIP, 2009.

[60] S.J. Pollock, S.V. Chasteen, M. Dubson, and K.K. Perkins. The use of concept tests and peer instruction in upper-division physics. In *AIP Conference Proceedings*, volume 1289, pages 261–264. AIP, 2010.

[61] 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(2):020114, 2017.

[62] 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, Aug 2016.

[63] 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. Phys. Educ. Res.*, 9:020121, Nov 2013.

[64] J. Docktor and K. Heller. Gender differences in both Force Concept Inventory and introductory physics performance. In *AIP Conference Proceedings*, volume 1064, pages 15–18. AIP Publishing, 2008.

[65] J. Margolis and A. Fisher. *Unlocking the Clubhouse: Women in Computing.* MIT Press, Cambridge, MA, 2003.

[66] 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.

[67] 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.

[68] B.C. Cunningham, K.M. Hoyer, and D. Sparks. *The Condition of STEM 2016.* ACT, Iowa City, IA, 2016.

[69] P.M. Sadler and R.H. Tai. Success in introductory college physics: The role of high school preparation. *Sci. Educ.*, 85:111–136, 2001.

[70] 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(6):847–876, 2007.

[71] T. Antimirova, A. Noack, and M. Milner-Bolotin. The effect of classroom diversity on conceptual learning in physics. In *AIP Conference Proceedings*, volume 1179, pages 77–80. AIP Publishing, 2009.

[72] L.E. Kost, S.J. Pollock, and N.D. Finkelstein. Characterizing the gender gap in introductory physics. *Phys. Rev. Phys. Educ. Res.*, 5:010101, Jan 2009.

[73] P.B. Kohl and H.V. Kuo. Introductory physics gender gaps: Pre-and post-studio transition. In *AIP Conference Proceedings*, volume 1179, pages 173–176. AIP, 2009.

[74] 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. Phys. Educ. Res.*, 6:020112, Sep 2010.

[75] 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(2):421, 2013.

[76] N.S. Cole. *The ETS Gender Study: How Females and Males Perform in Educational Settings*. Educational Testing Service, Princeton, NJ, 1997.

[77] 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, NY, 2007.

[78] D. Voyer and S.D. Voyer. Gender differences in scholastic achievement: A meta-analysis. *Psychol. Bull.*, 140(4):1174, 2014.

[79] 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. Phys. Educ. Res.*, 2(1):010101, 2006.

[80] 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. Phys. Educ. Res.*, 6:010106, May 2010.

[81] V.P. Coletta, J.A. Phillips, and J. Steinert. FCI normalized gain, scientific reasoning ability, thinking in physics, and gender effects. In *AIP Conference Proceedings*, volume 1413, pages 23–26. AIP, 2012.

[82] M.Y. Jaber. *Learning Curves: Theory, Models, and Applications*. CRC Press, Tayler & Francis Group, New York, NY, 2016.

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

[84] S.B. Hofer, T.D. Mrsic-Flogel, T. Bonhoeffer, and M. Hübener. Experience leaves a lasting structural trace in cortical circuits. *Nature*, 457(7227):313–317, 2009.

[85] E.C. Sayre and A.F. Heckler. Peaks and decays of student knowledge in an introductory E&M course. *Phys. Rev. Phys. Educ. Res.*, 5(1):013101, 2009.

[86] D.F. Halpern. *Sex Differences in Cognitive Abilities, 4th ed.* Psychology Press, Francis & Tayler Group, New York, NY, 2012.

[87] 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. *Arch. Sex. Behav.*, 39(4):990–997, 2010.

[88] 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(1):69–94, 2013.

[89] J.S. Hyde and M.C. Linn. Gender differences in verbal ability: A meta-analysis. *Psychol. Bull.*, 104(1):53–69, 1988.

[90] 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. *Arch. Sex. Behav.*, 36(2):235–249, 2007.

[91] S.A. Sorby. Developing 3D spatial skills for engineering students. *Aust. J. Eng. Educ.*, 13(1):1–11, 2007.

[92] D.I. Miller and D.F. Halpern. The new science of cognitive sex differences. *Trends Cogn. Sci.*, 18(1):37–45, 2014.

[93] X. Ma. A meta-analysis of the relationship between anxiety toward mathematics and achievement in mathematics. *Jour. Res. Math. Educ.*, 30:520–540, 1999.

[94] 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(1):103, 2010.

[95] R.A. Alvaro. *The effectiveness of a science therapy program upon science anxious undergraduates.* PhD thesis, Loyola University Chicago, 1978.

[96] J.V. Mallow. A science anxiety program. *Am. J. Phys.*, 46(8):862–862, 1978.

[97] J.V. Mallow and S.L. Greenburg. Science anxiety: Causes and remedies. *J. Coll. Sci. Teach.*, 11:356–358, 1982.

[98] 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(4):356–369, 2010.

[99] 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(4):435–446, 2004.

[100] K. Williams. Understanding communication anxiety and gender in physics. *J. Coll. Sci. Teach.*, 30(4):232, 2000.

[101] N. Hall and D. Webb. Instructor's support of student autonomy in an introductory physics course. *Phys. Rev. Phys. Educ. Res.*, 10(2):020116, 2014.

[102] 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(2):119–130, 2008.

[103] 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(4):609–624, 2008.

[104] 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.

[105] C.C. de Cohen and N. Deterding. Widening the net: National estimates of gender disparities in engineering. *J. of Eng. Educ.*, 98(3):211–226, 2009.

[106] 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.

[107] H.D. Nguyen and A. Ryan. Does stereotype threat affect test performance of minorities and women? A meta-analysis of experimental evidence. *J. Appl. Psychol.*, 93(6):1314, 2008.

[108] G. Stoet and D.C. Geary. Can stereotype threat explain the gender gap in mathematics performance and achievement? *Rev. Gen. Psychol.*, 16(1):93–102, 2012.

[109] 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(9):1132–1139, 2009.

[110] J.R. Shapiro and A.M. Williams. The role of stereotype threats in undermining girls' and women's performance and interest in STEM fields. *Sex Roles*, 66(3-4):175–183, 2012.

[111] 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(3):299–333, 2013.

[112] 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(3-4):153–166, 2012.

[113] 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(2):020115, 2016.

[114] A. Maries, N.I. Karim, and Singh C. The impact of stereotype threat on gender gap in introductory physics. In *AIP Conference Proceedings*, pages 256–259. AIP, 2017.

[115] Laura McCullough. Gender differences in student responses to physics conceptual questions based on question context. *ASQ Advancing the STEM Agenda in Education*, 2011.

[116] 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. In *AIP Conference Proceedings*, pages 103–106. AIP, 2001.

[117] R. D. Dietz, R. H. Pearson, M. R. Semak, and C. W. Willis. Gender bias in the Force Concept Inventory? In *AIP Conference Proceedings*, volume 1413, pages 171–174. AIP, 2012.

[118] 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. In *2011 American Educational Research Association Conference*, Washington, DC, 2011. American Education Research Association.

[119] D.J. Low and K.F. Wilson. Persistent gender gaps in first-year physics assessment questions. In *Proceedings of The Australian Conference on Science and Mathematics Education*, pages 118–124, 2015.

[120] 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.

[121] R.J. Beichner and J.M. Saul. Introduction to the SCALE-UP (Student-Centered Activities for Large Enrollment Undergraduate Programs) project. In *Invention and impact: Building excellence in undergraduate Science, Technology, Engineering and Mathematics(STEM) education*, pages 61–66, Washington, DC, 2003. American Association for the Advancement of Science.

[122] M. Lorenzo, C.H. Crouch, and E. Mazur. Reducing the gender gap in the physics classroom. *Am. J. Phys.*, 74(2):118–122, 2006.

[123] S.W. Brahmia, C. Henderson, M. Sabella, and L. Hsu. Improving learning for underrepresented groups in physics for engineering majors. In *AIP Conference Proceedings*, volume 1064, pages 7–10. AIP Publishing, 2008.

[124] 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. Phys. Educ. Res.*, 3(1):010107, 2007.

[125] 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. Phys. Educ. Res.*, 10(2):020101, 2014.

[126] N.I. Karim, A. Maries, and C. Singh. Do evidence-based active-engagement courses reduce the gender gap in introductory physics? *Eur. J. Phys.*, 39:1–31, 2018.

[127] G.J. Privitera. *Statistics for the Behavioral Sciences.* Sage, Los Angeles, CA, 2015.

[128] R. Nuzzo. Scientific method: Statistical errors. *Nature News*, 506(7487):150, 2014.

[129] Student. The probable error of a mean. *Biometrika*, pages 1–25, 1908.

[130] J. Cohen. A power primer. *Psychol. Bull.*, 112(1):155, 1992.

[131] J. Cohen. Things I have learned (so far). *Am. Psychol.*, 45(12):1304, 1990.

[132] J.A.C. Hattie. *Visible Learning: A Synthesis of Over 800 Meta-Analyses Relating to Achievement.* Routledge, Taylor & Francis Group, New York, NY, 2009.

[133] J. Neyman and E.S. Pearson. On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. *Biometrika*, pages 175–240, 1928.

[134] J. Neyman and E.S. Pearson. On the use and interpretation of certain test criteria for purposes of statistical inference: Part II. *Biometrika*, pages 263–294, 1928.

[135] O.J. Dunn. Multiple comparisons among means. *J. Am. Stat. Assoc.*, 56(293):52–64, 1961.

[136] C.M. Judd, G.H. McClelland, and C.S. Ryan. *Data Analysis: A Model Comparison Approach.* Routledge, New York, NY, 2011.

[137] J. Templin. Item response theory. *The Encyclopedia of Adulthood and Aging*, 2011.

[138] J.C. Nunnally and I.H. Bernstein. *Psychometric Theory, Third Edition.* McGraw-Hill, New York, NY, 1994.

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

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

[141] D.B. McCoach, R.K. Gable, and J.P. Madura. *Instrument Development in the Affective Domain: School and Corporate Applications 3rd ed.* Springer, New York, NY, 2013.

[142] 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(4):454–496, 2015.

[143] P.M. Sadler. Psychometric models of student conceptions in science: Reconciling qualitative studies and distractor-driven assessment instruments. *J. Res. Sci. Teach*, 35(3):265–296, 1998.

[144] R.S. Lindell. *Enhancing college students' understanding of lunar phases.* PhD thesis, University of Nebraska, Lincoln, NE, 2001.

[145] R.J. De Ayala. *The Theory and Practice of Item Response Theory.* Guilford Publications, New York, NY, 2013.

[146] C. DeMars. *Item Response Theory.* Oxford University Press, New York, NY, 2010.

[147] G. Rasch. Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests. 1960.

[148] R.D. Bock. Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37(1):29–51, 1972.

[149] W.M. Yen. Using simulation results to choose a latent trait model. *Appl. Psych. Meas.*, 5(2):245–262, 1981.

[150] S.P. Reise. A comparison of item- and person-fit methods of assessing model-data fit in IRT. *Appl. Psych. Meas.*, 14(2):127–137, 1990.

[151] P.W. Holland and D.T. Thayer. An alternate definition of the ETS delta scale of item difficulty. *ETS Research Report Series*, 1985(2), 1985.

[152] P.W. Holland and D.T. Thayer. Differential item performance and the Mantel-Haenszel procedure. In H. Wainer and H.I. Braun, editors, *Test validity*, pages 129–145. Lawrence Erlbaum Associates, New Jersey, 1988.

[153] B.E. Clauser and K.M. Mazor. Using statistical procedures to identify differentially functioning test items. *Educ. Meas-Issues Pra.*, 17(1):31–44, 1998.

[154] R. Zwick and K. Ercikan. Analysis of differential item functioning in the naep history assessment. *J. Educ. Meas.*, 26(1):55–66, 1989.

[155] J. Liu, D.J. Harris, and A. Schmidt. Statistical procedures used in college admissions testing. In *Handbook of Statistics. Vol. 26. Psychometrics*, pages 1057–1091. Elsevier, Amsterdam, 2007.

[156] F. Lord. *Applications of Item Response Theory to Practical Testing Problems.* Lawrence Erlbaum Associates, Hillsdale, NJ, 1980.

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

[158] C. Spearman. "general intelligence," objectively determined and measured. *Am. J. Psych.*, 15(2):201–292, 1904.

[159] L.R. Fabrigar, D.T. Wegener, R.C. MacCallum, and E.J. Strahan. Evaluating the use of exploratory factor analysis in psychological research. *Psych. Meth.*, 4(3):272, 1999.

[160] L.R. Fabrigar and D.T. Wegener. *Exploratory Factor Analysis.* Oxford University Press, New York, NY, 2011.

[161] B. Thompson. *Exploratory and Confirmatory Factor Analysis: Understanding Concepts and Applications.* American Psychological Association, Washington DC, 2004.

[162] T.A. Brown. *Confirmatory Factor Analysis for Applied Research.* Guilford Publications, New York, NY, 2014.

[163] R.L. Gorsuch. Exploratory factor analysis: Its role in item analysis. *J. Per. Assess.*, 68(3):532–560, 1997.

[164] S. Wold, K. Esbensen, and P. Geladi. Principal component analysis. *Chemotetr. Intell. Lab.*, 2(1-3):37–52, 1987.

[165] M.A. Pett, N.R. Lackey, and J.J. Sullivan. *Making Sense of Factor Analysis: The Use of Factor Analysis for Instrument Development in Health Care Research.* Sage, Thousand Oaks, CA, 2003.

[166] L. Hu and P.M. Bentler. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Struct. Equ. Modeling*, 6(1):1–55, 1999.

[167] 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. Equ. Modeling*, 11(3):320–341, 2004.

[168] R.E. Schumacker and R.G. Lomax. *A Beginner's Guide to Structural Equation Modeling.* Lawrence Erlbaum Associates, Inc., Mahwah, NJ, 2004.

[169] C. Henderson, M. Dancy, and M. Niewiadomska-Bugaj. Use of research-based instructional strategies in introductory physics: Where do faculty leave the innovation-decision process? *Phys. Rev. ST Phys. Educ. Res.*, 8:020104, Jul 2012.

[170] R.J. Beichner, J.M. Saul, D.S. Abbott, J.J. Morse, D. Deardorff, R.J. Allain, S.W. Bonham, M.H. Dancy, and J.S. Risley. The student-centered activities for large enrollment undergraduate programs (scale-up) project. *Research-based Reform of University Physics*, 1(1):2–39, 2007.

[171] J. Handelsman, D. Ebert-May, R. Beichner, P. Bruns, et al. Scientific teaching. *Science*, 304(5670):521, 2004.

[172] V.P. Coletta, J.A. Phillips, and J.J. Steinert. Interpreting force concept inventory scores: Normalized gain and SAT scores. *Phys. Rev. Phys. Educ. Res.*, 3:010106, May 2007.

[173] US News & World Report: Education. `https://premium.usnews.com/best-colleges`. Accessed 4/30/2017.

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

[175] 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(12):943–950, 2013.

[176] 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(12):1259–1268, 2002.

[177] 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(6008):1234–1237, 2010.

[178] 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. In *AIP Conference Proceedings*, volume 1413, pages 231–234. AIP Publishing, 2012.

[179] 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(1):30–38, 2013.

[180] E. Kim and S. Pak. Students do not overcome conceptual difficulties after solving 1000 traditional problems. *Am. J. Phys.*, 70(7):759–765, 2002.

[181] B.J. Zimmerman. Self-efficacy: An essential motive to learn. *Contemp. Educ. Psychol.*, 25(1):82–91, 2000.

[182] 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 Assessment*, 23(3):399–409, 2014.

[183] 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(1):79–122, 1994.

[184] M. Richardson, C. Abraham, and R. Bond. Psychological correlates of university students' academic performance: A systematic review and meta-analysis. *Psychol. Bull.*, 138(2):353–387, 2012.

[185] J. Price. The effect of instructor race and gender on student persistence in STEM fields. *Econ. Educ. Rev.*, 29(6):901–910, 2010.

[186] D. Huffman and P. Heller. What does the Force Concept Inventory actually measure? *Phys. Teach.*, 33:138, 1995.

[187] T.F. Scott, D. Schumayer, and A.R. Gray. Exploratory factor analysis of a Force Concept Inventory data set. *Phys. Rev. Phys. Educ. Res.*, 8(2):020105, 2012.

[188] 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, Jan 2017.

[189] 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(5):449–453, 2006.

[190] J. Wang and L. Bao. Analyzing Force Concept Inventory with Item Response Theory. *Am. J. Phys.*, 78(10):1064–1070, 2010.

[191] M. Planinic, L. Ivanjek, and A. Susac. Rasch model based analysis of the Force Concept Inventory. *Phys. Rev. Phys. Educ. Res.*, 6:010103, Mar 2010.

[192] T. F. Scott and D. Schumayer. Students' proficiency scores within multitrait item response theory. *Phys. Rev. Phys. Educ. Res.*, 11(2):020134, 2015.

[193] 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(2):020131, 2016.

[194] N.J. Dorans. ETS contributions to the quantitative assessment of item, test, and score fairness. *ETS Research Report Series*, 2013(2), 2013.

[195] ETS Standards for Quality and Fairness. `https://www.ets.org/s/about/pdf/standards.pdf`. Accessed 11/11/2017.

[196] 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.

[197] M. Zieky. Fairness review in assessment. In *Handbook of Test Development*, pages 359–376. Lawrence Erlbaum Associates, Mahwah, NJ, 2006.

[198] 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, Jan 2018.

[199] D. Hestenes and I. Halloun. Interpreting the Force Concept Inventory: A response to March 1995 critique by Huffman and Heller. *Phys. Teach.*, 33(8):502–502, 1995.

[200] P. Heller and D. Huffman. Interpreting the Force Concept Inventory: A reply to Hestenes and Halloun. *Phys. Teach.*, 33(8):503–503, 1995.

[201] G.A. Morris, N. Harshman, L. Branum-Martin, E. Mazur, T. Mzoughi, and S.D. Baker. An item response curves analysis of the Force Concept Inventory. *Am. J. Phys.*, 80(9):825–831, 2012.

[202] 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*, Washington, DC, 2011. American Education Research Association.

[203] N. Lasry, S. Rosenfield, H. Dedic, A. Dahan, and O. Reshef. The puzzling reliability of the Force Concept Inventory. *Am. J. Phys.*, 79(9):909–912, 2011.

[204] C. Henderson. Common concerns about the Force Concept Inventory. *Phys. Teach.*, 40(9):542–547, 2002.

[205] G. Novak, A. Gavrin, W. Christian, and E. Patterson. *Just-In-Time Teaching: Blending Active Learning with Web Technology.* Addison-Wesley, Upper Saddle River, NJ, 1st edition, 1999.

[206] S. DeVore, J. Stewart, and G. Stewart. Examining the effects of testwiseness in conceptual physics evaluations. *Phys. Rev. Phys. Educ. Res.*, 12:020138, 2016.

[207] J.S. Hyde, E. Fennema, and S.J. Lamon. Gender differences in mathematics performance: a meta-analysis. *Psychol. Bull.*, 107(2):139, 1990.

[208] *The American Heritage Dictionary of the English Language.* Houghton Mifflin Co., Boston, MA, 2000.

[209] J. Han, L. Bao, L. Chen, T. Cai, Y. Pi, S. Zhou, Y. Tu, and K. Koenig. Dividing the Force Concept Inventory into two equivalent half-length tests. *Phys. Rev. Phys. Educ. Res.*, 11:010112, May 2015.

[210] R.M. Talbot. Taking an item-level approach to measuring change with the Force and Motion Conceptual Evaluation: An application of item response theory. *Sch. Sci. Math.*, 113(7):356–365, 2013.

[211] 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. Phys. Educ. Res.*, 10(2):020114, 2014.

[212] R.K. Thornton, D. Kuhl, K. Cummings, and J. Marx. Comparing the Force and Motion Conceptual Evaluation and the Force Concept Inventory. *Phys. Rev. Phys. Educ. Res.*, 5(1):010105, 2009.

[213] T.I. Smith and M.C. Wittmann. Applying a resources framework to analysis of the Force and Motion Conceptual Evaluation. *Phys. Rev. Phys. Educ. Res.*, 4(2):020101, 2008.

[214] 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. Phys. Educ. Res.*, 10(2):020102, 2014.

[215] S. Ramlo. Validity and reliability of the Force and Motion Conceptual Evaluation. *Am. J. Phys.*, 76(9):882–886, 2008.

[216] The Carnegie Classification of Institutions of Higher Education. `http://carnegieclassifications.iu.edu/`. Accessed 9/21/2017.

[217] R.K. Thornton. Conceptual dynamics: Following changing student views of force and motion. In *AIP Conference Proceedings*, volume 399, pages 241–266. AIP, 1997.

[218] D.E. Meltzer. Analysis of shifts in students? reasoning regarding electric field and potential concepts. In *AIP Conference Proceedings*, volume 883, pages 177–180. AIP, 2007.

[219] J. Leppävirta. The effect of naïve ideas on students? reasoning about electricity and magnetism. *Res. Sci. Educ.*, 42(4):753–767, 2012.

[220] R. Scherr. Never mind the gap: Gender-related research in *Physical Review Physics Education Research*, 2005–2016. *Phys. Rev. Phys. Educ. Res.*, 12(2):020003, 2016.

[221] National Science Foundation, Nation Center for Science, and Engineering Statistics. *Science and Engineering Degrees, by Race/Ethnicity of Recipients: 2002-2012. Detailed Statistical Tables NSF 15-321.* National Science Foundation, Arlington, VA, 2015.

[222] L. Perna, V. Lundy-Wagner, N.D. Drezner, M. Gasman, S. Yoon, E. Bose, and S. Gary. The contribution of hbcus to the preparation of african american women for stem careers: A case study. *Res. High. Educ.*, 50(1):1–23, 2009.

[223] V.A. Lotkowski, S.B. Robbins, and R.J. Noeth. *The Role of Academic and Non-Academic Factors in Improving College Retention. ACT Policy Report.* ACT, Inc, 2004.

[224] J.M. Braxton, J.F. Milem, and A.S. Sullivan. The influence of active learning on the college student departure process: Toward a revision of tinto's theory. *J. High. Educ.*, 71(5):569–590, 2000.

[225] T.T. Ishitani. How do transfers survive after transfer shock? a longitudinal study of transfer student departure at a four-year institution. *Res. High. Educ.*, 49(5):403–419, 2008.

[226] B. Toven-Lindsey, M. Levis-Fitzgerald, P.H. Barber, and T. Hasson. Increasing persistence in undergraduate science majors: A model for institutional support of underrepresented students. *CBE-Life Sci. Educ.*, 14(2):ar12, 2015.

[227] D.F. Carter. Key issues in the persistence of underrepresented minority students. *New Dir. Inst. Res.*, 2006(130):33–46, 2006.

[228] S. Hurtado, C.B. Newman, M.C. Tran, and M.J. Chang. Improving the rate of success for underrepresented racial minorities in stem fields: Insights from a national project. *New Dir. Inst. Res.*, 2010(148):5–15, 2010.