1-2018

Gender fairness within the Force Concept Inventory

Adrienne L. Traxler
Wright State University, adrienne.traxler@wright.edu

Rachel Henderson

John Stewart

Gay Stewart

Alexis Papak

See next page for additional authors

Editors' Suggestion

Gender fairness within the Force Concept Inventory

Adrienne Traxler,[1,*] Rachel Henderson,[2] John Stewart,[2] Gay Stewart,[2]
Alexis Papak,[3] and Rebecca Lindell[4]

[1]*Department of Physics, Wright State University, Dayton, Ohio 45435, USA*
[2]*Department of Physics and Astronomy, West Virginia University, Morgantown, West Virginia 26506, USA*
[3]*Department of Physics, University of Maryland, College Park, Maryland 20742, USA*
[4]*Tiliadal STEM Education Solutions, Lafayette, Indiana 47901, USA*

(Received 1 September 2017; published 18 January 2018)

Research on the test structure of the Force Concept Inventory (FCI) has largely ignored gender, and research on FCI gender effects (often reported as "gender gaps") has seldom interrogated the structure of the test. These rarely crossed streams of research leave open the possibility that the FCI may not be structurally valid across genders, particularly since many reported results come from calculus-based courses where 75% or more of the students are men. We examine the FCI considering both psychometrics and gender disaggregation (while acknowledging this as a binary simplification), and find several problematic questions whose removal decreases the apparent gender gap. We analyze three samples (total $N_{\text{pre}} = 5391$, $N_{\text{post}} = 5769$) looking for gender asymmetries using classical test theory, item response theory, and differential item functioning. The combination of these methods highlights six items that appear substantially unfair to women and two items biased in favor of women. No single physical concept or prior experience unifies these questions, but they are broadly consistent with problematic items identified in previous research. Removing all significantly gender-unfair items halves the gender gap in the main sample in this study. We recommend that instructors using the FCI report the reduced-instrument score as well as the 30-item score, and that credit or other benefits to students not be assigned using the biased items.

DOI: 10.1103/PhysRevPhysEducRes.14.010103

I. INTRODUCTION

The Force Concept Inventory (FCI) [1] has been studied using tools such as factor analysis [2,3], item response theory [4,5], and network analysis [6]. Though these investigations have probed the structure and validity of the test, they have primarily treated student data as a single undifferentiated sample and have not studied gender effects. A largely separate branch of research has explored gender differences in scores on the FCI and other conceptual inventories [7,8]. These studies have documented a ubiquitous advantage for men on pretest questions, which often persists to the post-test. Proposed explanations range from differences in preparation, to instructional method (when examining gains), to sociocultural factors such as stereotype threat. With some exceptions, the literature on test construction largely ignores gender effects, and the literature on gender effects focuses on total score and takes the integrity of the instrument as a given. Because a great deal of FCI data are collected from calculus-based courses where 75% or more of the students are male, it remains an open question whether gender-blind validations of the FCI for "all students" are in fact applicable to all, or whether poorly functioning items for women might be hidden in the unbalanced sample.

In this paper, gender fairness is explored in three samples of FCI pretest and post-test data (total $N_{\text{pre}} = 5391$, $N_{\text{post}} = 5769$). We employ classical test theory, item response theory, and differential item functioning analysis to determine if FCI items are equally fair for men and women. We explore two dimensions of fairness: item fairness and test construction fairness. An item is defined as being "fair" if men and women of equal ability have the same chance of answering the item correctly. An instrument is defined as having "test construction fairness" if the instrument and items within the instrument have similar performance on test evaluation metrics for men and women. An evaluation of fairness is a crucial step in the test development process [9–11].

We acknowledge that a binary view of gender in physics education is at best a first-order model, simplifying a wide range of sociocultural factors and nuanced gender identities

*Corresponding author.
adrienne.traxler@wright.edu

into two categories [8]. Nonetheless, this model has been the basis for reporting many score differences on standardized instruments such as the FCI. This work focuses on fairness for men and women; future research should examine fairness for other marginalized groups.

In the remainder of this introduction we summarize "gender gap" findings for the FCI, note the most popular student-based causes that have been proposed, and describe our psychometric framework for analyzing the instrument. This framework draws in part on that of Jorion *et al.* [12], which maps a process for validating conceptual inventories, but which we expand by incorporating item fairness as part of the process. This introduction will discuss many items within the FCI; for readers unfamiliar with the instrument a description of the instrument and a summary of the items is provided in Sec. II A. Like many technical fields, psychometrics has its own set or terms and definitions that may differ from the common definition. We will introduce definitions for the terms used in this work as they are encountered. For a careful set of definitions of psychometric terms and suggested practice see *Standards for Educational and Psychological Testing* [13]; for concise definitions see the glossary in the same volume.

A. Gender gap investigations of the FCI

The FCI has been used for measuring student conceptual gains in introductory mechanics for nearly 25 years. For more than half that period, published studies have documented an apparent gender difference in item responses, overall scores, and instructional gains. Madsen, McKagan, and Sayre provide an overview of the research into the gender gap in conceptual instruments used in physics education research (PER) [7]. On average, male students outperform female students by 13% on pretests and 12% on post-tests of conceptual mechanics instruments, the FCI and the Force and Motion Conceptual Evaluation (FMCE) [14]. Men also outperformed women by 8.5% on post-tests of electricity and magnetism instruments. This effect is nearly universal with only one of the seventeen studies showing a female advantage on the post-test.

Most of the studies reported in the Madsen, McKagan, and Sayre review follow common educational research practice which locates the source of the gap within the students. Suggested influences in gender-based performance differences include documented differences in male or female high school physics class election [15–17] and the effect of these differences on college physics grades [18,19]. A large body of research also shows differences in academic course grades [20,21] and performance on cognitive tests [22–25] with women scoring higher on verbal reasoning and men scoring higher on spatial reasoning. Physics-specific variations on this research have examined declared major, years of high school calculus, and correlations with the Lawson test of scientific reasoning or other standardized tests as a proxy for broader cognitive abilities (see Madsen, McKagan, and Sayre [7], Table I for summary).

Many psychological factors have also been investigated to explain gender differences such as mathematics anxiety [26,27], science anxiety [28–30], and stereotype threat [31]. In physics education research, psychological explanations have included self-efficacy, endorsement of gender stereotypes, or attitudes toward physics ([7], Table I). It is much harder to find studies that investigate gender bias in university physics learning environments, though work in science education has linked such bias to the greater attrition of women from many STEM fields [32,33].

Results show decreased gender gaps in classrooms using some active-learning curricula [34–36] which may provide an avenue to reduce attrition. However, these results have been inconsistent; other results show no reduction of the gender gap in classrooms using active engagement [37–39]. A great deal of work remains to be done in this area, and it is likely to require detailed qualitative data collection and analyses that are substantially more time consuming to conduct than pre- and post conceptual inventory measures.

A third possible source of conceptual inventory gender gaps, that of bias in the test questions, can be analyzed by later researchers even if it is not considered during instrument design. For the FCI, several studies have highlighted items using psychometric analysis that appear to function differently for students of different genders. These findings have typically not received as much attention as more student-centered explanations for performance differences. We will highlight these studies in the following sections that expand on the psychometric framework, and return to them in our discussion of results.

The FCI continues to be used as a diagnostic of student understanding, and in many cases to assign course credit, despite a trail of evidence of gender bias. For an overview of research on the gender gap in physics conceptual inventories see Madsen, McKagan, and Sayre [7]. For a more recent summary of research into possible explanations of male and female performance differences in physics see Henderson *et al.* [40]. For a more general discussion of gender in physics see Traxler *et al.* [8]. For an overview of gender disparities in STEM see Eddy and Brownell [41].

B. Validity framework

Item analysis is usually performed at the beginning of the test validation process to identify items which may be a threat to the reliability of the instrument. An instrument is reliable if multiple applications of the instrument in similar testing conditions yield similar results [13]. An instrument with poor reliability cannot have strong validity. An instrument is valid if it accurately measures the constructs it was designed to measure [13]. A review of the literature did not identify any published work formally performing an item analysis for the FCI. We use the framework of Jorion

et al. [12], developed for evaluating the validity of conceptual inventories in engineering education. Their framework collects some standard methods of item analysis used in classical test theory (CTT) and item response theory (IRT) [42]. First, they use thresholds for item difficulty and discrimination from CTT to flag potentially poorly constructed items. Items with poor performance on some psychometric measures are called "problematic." The item characteristic curves from IRT are then examined to determine if some items were problematic within the IRT model. Cronbach's alpha and interitem correlations are then calculated to identify items that may have reliability problems. The factor structure of the instrument is then compared with the factor structure published by the instrument's creators.

The framework does not address the issue of using a common instrument for both pretest and post-test with student populations of varying academic capability. Furthermore, it does not evaluate item fairness, a critical oversight for conceptual instruments used in class environments where some populations of students are seriously underrepresented. We adopt the CTT and IRT measures used by Jorion *et al.* We extend the framework to include item fairness analysis using differential item functioning as discussed below.

The validity and reliability checks in the Jorion framework should be performed at the beginning of instrument development. These methods are far from complete. Once a set of reliable and fair items is identified, additional analysis is required to demonstrate these items measure the intended constructs. An impressive array of evidence attests to both the face and criterion validity of the FCI and its test-retest reliability for gender aggregated samples [43,44].

C. Difficulty and discrimination

Establishing the validity of an instrument is a multifacted process that must be repeated for all populations of interest. A first step considers two basic tools of item analysis, difficulty (P) and discrimination (D). CTT suggests well-performing items should have $0.2 < P < 0.8$ and $D > 0.2$ [12]. For a review of CTT and IRT see Ding and Beichner [45], or see Sec. II and Supplemental Material [46] for details of their use in this paper.

While many studies employ the FCI, few report item level statistics. Wang and Bao calculated CTT difficulty and discrimination parameters for the FCI pretest of 2800 students at a large university in the U.S. [4]. Five of the items had difficulty parameters outside of the desired range (items 1, 6, and 12 with $P > 0.8$ and items 17 and 26 with $P < 0.2$), with none having discrimination less than 0.2. Morris *et al.* reported the item averages of 4500 students pooling data from multiple institutions and reported FCI items 5, 17, and 26 with $P < 0.2$, but no items with $P > 0.8$ [5]. Osborn Popp, Meltzer, and Megowan-Romanowicz reported FCI item level scores for 4775 high school students. For male students, items 1, 6, and 16 had $P > 0.8$; for female students item 26 had $P < 0.2$ [47].

IRT also estimates difficulty and discrimination and can be used to explore validity and fairness. Many different IRT models have been applied to the FCI [4,47–49]. Of these studies, only Wang and Bao [4] reported the item characteristic curves which show how well the data fit the IRT model; none of their curves showed the dramatic departures from fit reported for some of the engineering conceptual inventories examined by Jorion *et al.* [12], indicating that the items in the FCI are generally performing properly. Only Popp *et al.* [47] reported results disaggregated by gender; these results are describe in Sec. I E. IRT models are discussed in more detail in the Supplemental Material [46].

D. Reliability

CTT also provides measures of instrument reliability. Lasry *et al.* assessed the overall reliability of the FCI by measuring both test-retest performance and internal consistency [43]. Their study reported the Kuder-Richardson reliability coefficient (KR-20), which had the value 0.9 for the initial application of the FCI and 0.865 combining the initial test and a retest given one week later. The KR-20 statistic is equivalent to Cronbach's alpha (used in the Jorion *et al.* framework) for dichotomous items such as those on the FCI. Values of KR-20 greater than 0.7 represent acceptable internal consistency [50]. Henderson [44] also examined test-retest reliability between the FCI as a graded posttest and as an ungraded quiz given the following semester; excellent test-retest reliability was measured in a sample of 500 university students. The FCI has also been compared with an alternate test of conceptual knowledge of mechanics, the FMCE [14]; a high correlation of overall test scores, $r = 0.78$, was demonstrated [51].

One can also examine subscale reliability, whether subgroups of questions thought to measure the same construct vary together. Factor analysis is often employed to identify these subgroups. The FCI authors proposed a division of the instrument into subcategories [1], but exploratory factor analysis failed to reproduce this division [2,52,53]. More recent analyses have resolved an alternate factor structure [3,48]; however, replication studies are needed to determine if these structures are robust. Because there is not yet a consensus on the FCI factor structure, we did not perform a confirmatory factor analysis.

E. Item fairness

Test and item fairness is a complex and sometimes contentious topic [10]. In this work, differential item functioning (DIF) analysis is used to explore whether the scores on individual FCI items are fair. This work will employ a narrow definition of a fair item as an item with

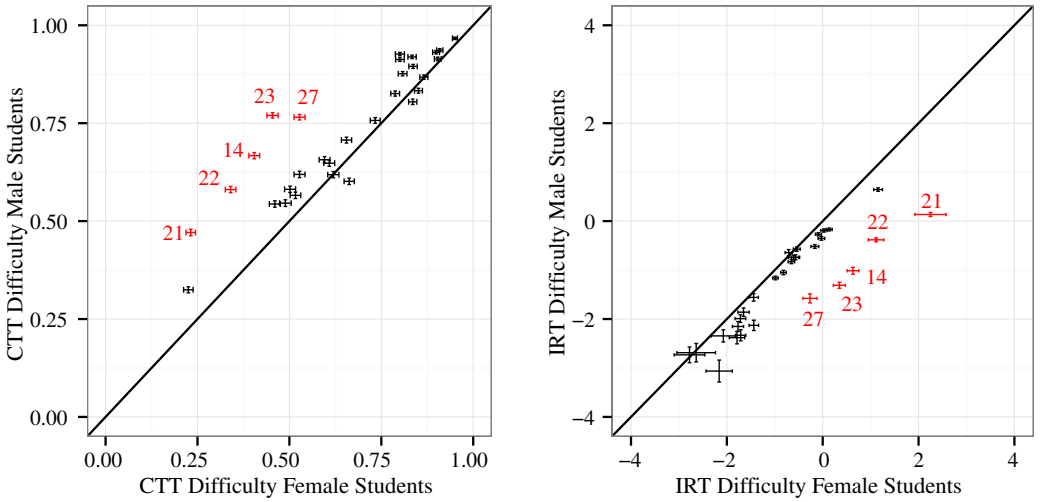

FIG. 1. CTT and IRT post-test results for Sample 1. Items 14, 21, 22, 23, and 27 are marked in red and labeled. A line of slope one is drawn to allow comparison of male and female difficulty. Error bars represent 1 standard deviation in each direction.

null DIF; that is, the score on the item is fair to multiple groups of participants if members of each group with the same ability (measured by overall FCI score) generate similar outcomes on the item. Fairness is identified as a key element in test development by the Educational Testing Service (ETS) and DIF analysis as a key step in evaluating fairness [9]. For an overview of item fairness and its relation to DIF see Dorans [11]. For a review of the complex issue of test and item fairness see Zieky [10].

DIF analysis provides statistics to assess the score fairness of items for subgroups of participants who have different abilities. Many DIF statistics have been constructed; this work uses the Mantel-Haenszel (MH) statistic [54,55], and Lord's statistic, L, an IRT alternative.

Dietz *et al.* used the MH statistic to evaluate DIF in an approximately gender-balanced sample of 520 students and found FCI items 4 and 9 were significantly biased against men and item 23 biased against women ($p < 0.005$), all with large DIF [56]. They also presented plots similar to Figs. 1 and 2 (Sec. III B). Their results showed many items were substantially unfair to women; however, error bars were not presented so it was difficult to assess whether these effects were the result of sample variance. They acknowledge their results were limited by sample size. While challenging to interpret because the data were plotted on a logarithmic scale, if the averages remained stable as sample size was increased, many items would exhibit small to moderate or large DIF including items 6,

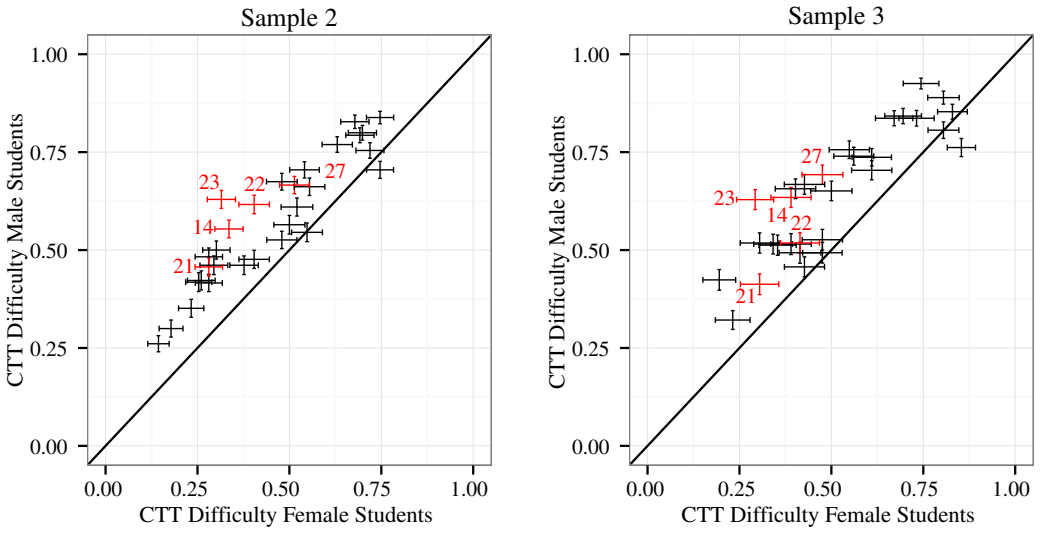

FIG. 2. CTT post-test difficulty results for Sample 2 and 3. Items 14, 21, 22, 23, and 27 are marked in red. A line of slope one is drawn to allow comparison of male and female difficulty.

12, 14, and 27, which are identified as problematic in our results below.

Osborn Popp, Meltzer, and Megowan-Romanowicz investigated DIF in the FCI in a sample of 4775 high school students who had completed a high school physics course using Modeling Instruction [47]. They found 14 items with significant DIF, where a Bonferroni correction had been applied to correct the p value for the number of statistical tests performed. Their statistic can be converted to the ETS effect size by multiplying by 2.35. With this conversion, for the significant items, item 23 had large DIF while items 4, 6, 9, 14, 15, and 29 had small to moderate DIF.

McCullough and Meltzer [57] compared the performance of 222 algebra-based physics students on the original FCI and a version where each problem was modified to have a context thought to be more stereotypically familiar to women. They found significant differences in performance on items 14, 22, 23, and 29. Using a similar methodology applied to nonphysics students, McCullough [58] showed female performance did not change while male performance decreased on the FCI modified to stereotypically female contexts.

As such, there is substantial but inconsistent support for the existence of gender unfair items in the FCI. This study seeks to answer the following 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?

II. METHODS

This study reports results from CTT, IRT, and DIF analyses. Table I summarizes the measures and their typical values.

A. The 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 [1]. 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 work uses the revised instrument published with Mazur [59] which is available at PhysPort [60].

This and other studies have identified items which may be unfair to either men or women; we provide a brief description of the most consistently identified items. 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.

TABLE I. Summary of item statistics, goodness-of-fit measures, and effect sizes reported in this study.

Measure	Description	Usage and range notes
		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 [50].
ϕ	Pearson correlation	Effect size of difference between P_F and P_M: 0.1 small, 0.3 medium, 0.5 large
		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)
d	Cohen's d	Gender difference in calculated difficulty; 0.2 small, 0.5 medium, 0.8 large
V	Cramer's V	Goodness of fit; 0.1 small misfit, 0.3 medium, 0.5 large
		DIF
$\Delta\alpha_{MH}$	Mantel-Haenszel	$\lvert\Delta\alpha_{\mathrm{MH}}\rvert < 1$, 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

B. Samples

This study employs three data sets collected at four U.S. universities. Racial or ethnic demographics were not available for individual students in the data but are reported at the university level.

1. Sample 1

Sample 1 was collected from a large, southern land-grant university enrolling approximately 25 000 students. In 2012, university demographics by race or 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 [61]. The sample was collected from the Spring 2002 semester to the Fall 2012 semester. The data set contains 4509 complete pretest responses (22.8% female) and 4716 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-min lectures and two 2-h 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 II) 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 data sets.

2. 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 or 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 [59], to heavily interactive classes using Peer Instruction, Just-in-Time Teaching [62], 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 [61]. The other institution had a range of SAT scores (25th percentile to 75 percentile) of 890–1130, which is equivalent to the 18–25 range of ACT scores [61].

TABLE II. 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.

		Male students		Female students		
	N	N	(M ± SD)%	N	(M ± 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

3. 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 or 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 [63]. 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 data set 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 [61].

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 Samples 2 and 3 combined, the aggregated data set would largely produce the same results as Sample 1.

C. CTT analyses

In CTT [42], item difficulty P is defined as the proportion of participants that answer an item correctly for a given population (thus, higher values indicate easier items). Item discrimination D is defined as [42]

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

where P_u is the proportion of participants in the top 27% of the total score distribution answering the question correctly and P_l is the proportion of participants in the bottom 27% answering the item correctly. An item with low or negative discrimination would be answered correctly by a substantial percentage of low-scoring students and incorrectly by high-scoring students, and might be poorly phrased or mostly answered by guessing.

For distractor-driven instruments, where the incorrect responses are drawn from attractive alternate ideas, an item is judged to be appropriate if its discrimination is above 0.2 [12,64,65]. In addition, items should not be either too difficult or too easy, resulting in difficulty cutoffs below 0.2 and above 0.8 [12]. Items that fall outside these cutoffs are classified as problematic and would normally be considered for elimination during the test construction process. In addition, we calculated Cronbach's alpha for reliability and checked interitem correlations; this analysis is presented in the Supplemental Material [46].

D. IRT analyses

CTT treats each item independently when calculating difficulty, ignoring the repeated-measures nature of an examination containing multiple items. CTT, therefore, ignores correlations resulting from the differing abilities of test takers. Item response theory explicitly models the effect of differing abilities by introducing a latent trait θ_i, which varies by participant i and is related to the probability that the participant answers a question correctly independent of the item.

IRT is an expansive topic with models for many testing situations [66]. The model most closely related to CTT is called the 2PL model, or two-parameter logistic model. This model assumes that each item j has a discrimination a_j and a difficulty b_j. The probability π_{ij} that participant i answers item j correctly is given by the logistic function:

$$\pi_{ij} = \frac{\exp[a_j(\theta_i - b_j)]}{1 + \exp[a_j(\theta_i - b_j)]}. \tag{2}$$

From Eq. (2), the probability of any set of item responses can be calculated and maximum likelihood estimation techniques employed to fit the parameters a_j, b_j, and θ_i. For a discussion of alternatives to the 2PL model, and goodness-of-fit tests for IRT, see the Supplemental Material [46].

E. DIF analyses

Differential item functioning will be measured with the Mantel-Haenszel statistic and Lord's statistic. The Mantel-Haenszel (MH) statistic [54,55], α_{MH}, is computed as a common odds ratio for an item using the total score on the instrument to form strata; thus, it pools the odds of a focal group (female students in this study) to answer correctly compared to a reference group (male students) for each level of ability, measured by overall score. Negative values indicate an advantage to male students, positive values an advantage to female students. An effect size can be constructed through a logarithmic transformation of the statistic $\Delta\alpha_{\mathrm{MH}} = -2.35\ln(\alpha_{\mathrm{MH}})$ [67]. This effect size measure was adopted by the Educational Testing Service (ETS) and is called the ETS delta scale; it has been in use for over 25 years [68]. The ETS classifies $|\Delta\alpha_{\mathrm{MH}}| < 1$ as negligible DIF, $1 \leq |\Delta\alpha_{\mathrm{MH}}| < 1.5$ as small to moderate DIF, and $|\Delta\alpha_{\mathrm{MH}}| \geq 1.5$ as large DIF. Lord's statistic L characterizes DIF in IRT and is scaled to the same effect size range. For details on Lord's statistic and more on the MH statistic, see the Supplemental Material [46].

F. Bonferroni correction

This work reports the statistical significance of many quantities and thus performs many statistical tests. To correct for the inflation of type I error rate, a Bonferroni correction was applied to each set of analyses by dividing the critical p values by the number of tests performed. For example, for the ϕ coefficient in Table III, $p = 0.05$ was changed to $p = 0.05/30 = 0.0017$ to correct for the 30 statistical tests performed for the 30 FCI items.

All statistical calculations were performed using the "R" statistical software [69]. IRT calculations were performed using the R package "ltm" [70], and DIF calculations used the R package "difR" [71].

III. RESULTS

Table II presents overall FCI pretest and post-test averages for the three samples. Significant gender differences ($p < 0.001$) were measured for all applications of the FCI, with Cohen's d [72] 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 [20] and superior male performance on standardized quantitative instruments [23,25], 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.

A. 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 III and difficulty plotted in Fig. 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 Samples 2 and 3 are plotted in Fig. 2. Male and female students were investigated separately. The standard deviations for the CTT parameters were calculated by bootstrapping using 1000 subsamples. Table IV presents the problematic items identified in the FCI for each sample. Critically, many of the questions flagged for female students in Table IV 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 IV 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

TABLE III. 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" $p < 0.000\,33$, and "c" $p < 0.000\,033$.

	Classical test theory					Item response theory							DIF	
No.	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	0.97 ± 0.00	0.95 ± 0.01	0.10 ± 0.01	0.13 ± 0.02	0.04	-2.71 ± 0.16	-2.78 ± 0.32	1.63 ± 0.14	1.30 ± 0.21	0.01	0.02	0.03	0.33	0.10
2	0.66 ± 0.01	0.60 ± 0.02	0.56 ± 0.02	0.44 ± 0.04	0.05^b	-0.74 ± 0.05	-0.61 ± 0.11	1.09 ± 0.06	0.73 ± 0.08	0.05	0.02	0.03	0.44	0.50
3	0.91 ± 0.00	0.90 ± 0.01	0.22 ± 0.01	0.25 ± 0.03	0.01	-2.15 ± 0.10	-1.77 ± 0.12	1.42 ± 0.09	1.89 ± 0.20	0.07	0.02	0.04	1.17^b	0.84
4	0.62 ± 0.01	0.62 ± 0.01	0.59 ± 0.02	0.57 ± 0.03	0.00	-0.57 ± 0.04	-0.54 ± 0.07	1.05 ± 0.05	1.19 ± 0.11	0.01	0.02	0.05	1.28^c	1.26^c
5	0.58 ± 0.01	0.50 ± 0.01	0.63 ± 0.02	0.65 ± 0.03	0.07^c	-0.35 ± 0.04	-0.03 ± 0.07	1.24 ± 0.06	1.08 ± 0.10	0.15^c	0.02^a	0.06^a	0.50	0.35
6	0.91 ± 0.00	0.80 ± 0.01	0.22 ± 0.01	0.34 ± 0.03	0.15^c	-2.34 ± 0.12	-2.07 ± 0.25	1.23 ± 0.09	0.75 ± 0.10	0.04	0.02	0.03	-1.43^c	-1.34^c
7	0.88 ± 0.01	0.81 ± 0.01	0.22 ± 0.02	0.28 ± 0.03	0.08^c	-2.69 ± 0.19	-2.64 ± 0.40	0.81 ± 0.07	0.58 ± 0.10	0.00	0.02	0.03	-0.45	-0.21
8	0.89 ± 0.01	0.84 ± 0.01	0.26 ± 0.01	0.37 ± 0.03	0.08^c	-2.13 ± 0.11	-1.44 ± 0.10	1.26 ± 0.08	1.62 ± 0.16	0.12^c	0.02	0.06^a	-0.14	-0.17
9	0.80 ± 0.01	0.84 ± 0.01	0.38 ± 0.02	0.40 ± 0.03	0.03	-1.56 ± 0.08	-1.44 ± 0.10	1.12 ± 0.06	1.59 ± 0.15	0.03	0.03^c	0.06^a	1.89^c	1.76^c
10	0.93 ± 0.00	0.90 ± 0.01	0.21 ± 0.01	0.28 ± 0.03	0.05^a	-1.99 ± 0.08	-1.72 ± 0.11	1.95 ± 0.13	1.92 ± 0.21	0.06	0.02	0.05	0.39	0.08
11	0.76 ± 0.01	0.73 ± 0.01	0.53 ± 0.02	0.63 ± 0.03	0.02	-1.05 ± 0.04	-0.82 ± 0.06	1.53 ± 0.07	2.15 ± 0.18	0.09^a	0.03^c	0.06^b	1.31^c	0.87^b
12	0.93 ± 0.00	0.80 ± 0.01	0.16 ± 0.01	0.31 ± 0.03	0.17^c	-3.06 ± 0.22	-2.16 ± 0.27	0.94 ± 0.08	0.71 ± 0.10	0.07	0.02	0.02	-1.97^c	-1.84^c
13	0.83 ± 0.01	0.79 ± 0.01	0.50 ± 0.02	0.57 ± 0.03	0.04	-1.16 ± 0.04	-0.99 ± 0.06	2.39 ± 0.12	2.51 ± 0.22	0.08	0.02	0.05^a	1.22^c	0.53
14	0.67 ± 0.01	0.40 ± 0.01	0.46 ± 0.02	0.44 ± 0.04	0.23^c	-1.01 ± 0.07	0.63 ± 0.12	0.78 ± 0.05	0.66 ± 0.08	0.39^c	0.02^a	0.06^b	-1.97^c	-1.84^c
15	0.60 ± 0.01	0.66 ± 0.02	0.45 ± 0.02	0.54 ± 0.04	0.05^b	-0.64 ± 0.06	-0.71 ± 0.07	0.72 ± 0.05	1.28 ± 0.11	0.02	0.05^c	0.08^c	1.77^c	2.00^c
16	0.94 ± 0.00	0.91 ± 0.01	0.17 ± 0.01	0.28 ± 0.03	0.04	-2.33 ± 0.11	-1.71 ± 0.11	1.51 ± 0.11	2.15 ± 0.24	0.10^b	0.02	0.05	0.36	0.17
17	0.55 ± 0.01	0.49 ± 0.02	0.67 ± 0.02	0.62 ± 0.03	0.05^a	-0.19 ± 0.03	0.03 ± 0.07	1.42 ± 0.06	1.19 ± 0.10	0.11^a	0.02	0.05	0.84^c	0.62
18	0.57 ± 0.01	0.52 ± 0.02	0.68 ± 0.02	0.69 ± 0.03	0.04	-0.27 ± 0.03	-0.09 ± 0.06	1.44 ± 0.06	1.27 ± 0.11	0.09	0.02	0.05	1.04^c	0.70^a
19	0.87 ± 0.01	0.87 ± 0.01	0.29 ± 0.02	0.33 ± 0.03	0.00	-1.86 ± 0.09	-1.65 ± 0.12	1.28 ± 0.08	1.56 ± 0.16	0.04	0.02	0.06^b	1.35^c	1.14^c
20	0.65 ± 0.01	0.61 ± 0.01	0.53 ± 0.02	0.55 ± 0.03	0.03	-0.74 ± 0.05	-0.57 ± 0.09	1.00 ± 0.05	0.95 ± 0.09	0.06	0.02	0.04	0.75^b	0.77^b
21	0.47 ± 0.01	0.23 ± 0.01	0.60 ± 0.02	0.29 ± 0.04	0.20^c	0.14 ± 0.04	2.25 ± 0.33	0.99 ± 0.05	0.57 ± 0.08	0.38^c	0.04^c	0.05	-1.86^c	-1.77^c
22	0.58 ± 0.01	0.34 ± 0.01	0.60 ± 0.02	0.42 ± 0.04	0.20^c	-0.38 ± 0.04	1.11 ± 0.16	1.08 ± 0.05	0.64 ± 0.08	0.45^c	0.03^c	0.07^c	-1.61^c	-1.56^c
23	0.77 ± 0.01	0.45 ± 0.02	0.45 ± 0.02	0.43 ± 0.04	0.29^c	-1.31 ± 0.06	0.35 ± 0.13	1.15 ± 0.06	0.55 ± 0.08	0.43^c	0.02	0.03	-2.70^c	-2.71^c
24	0.92 ± 0.00	0.83 ± 0.01	0.20 ± 0.01	0.32 ± 0.03	0.12^c	-2.38 ± 0.13	-1.79 ± 0.16	1.26 ± 0.09	1.10 ± 0.12	0.08	0.02	0.04	-0.94^b	-0.98^b
25	0.54 ± 0.01	0.46 ± 0.01	0.74 ± 0.02	0.66 ± 0.03	0.07^c	-0.17 ± 0.03	0.14 ± 0.06	1.72 ± 0.08	1.31 ± 0.11	0.17^c	0.03^c	0.06^a	0.70^a	0.32
26	0.32 ± 0.01	0.23 ± 0.01	0.66 ± 0.02	0.51 ± 0.04	0.09^c	0.64 ± 0.04	1.15 ± 0.09	1.65 ± 0.08	1.44 ± 0.13	0.22^c	0.03^c	0.05	0.40	-0.08
27	0.77 ± 0.01	0.53 ± 0.02	0.38 ± 0.02	0.37 ± 0.04	0.22^c	-1.58 ± 0.09	-0.27 ± 0.15	0.86 ± 0.05	0.45 ± 0.07	0.24^c	0.02^a	0.06^a	-1.87^c	-1.80^c
28	0.71 ± 0.01	0.66 ± 0.01	0.63 ± 0.02	0.62 ± 0.03	0.05^a	-0.83 ± 0.04	-0.65 ± 0.07	1.50 ± 0.07	1.37 ± 0.12	0.08	0.02^a	0.05	0.83^b	0.56
29	0.83 ± 0.01	0.85 ± 0.01	0.09 ± 0.02	0.14 ± 0.03	0.02	-18.4 ± 10	-5.24 ± 1.47	0.09 ± 0.05	0.34 ± 0.10	0.02	0.03^c	0.04	0.64	1.55^c
30	0.62 ± 0.01	0.53 ± 0.01	0.59 ± 0.02	0.55 ± 0.04	0.08^c	-0.52 ± 0.04	-0.16 ± 0.08	1.24 ± 0.06	0.86 ± 0.09	0.15^b	0.02	0.03	0.19	0.18

TABLE IV. CTT problematic items with $P < 0.2$, $P > 0.8$, or $D < 0.2$.

Gender	Pre or 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

items in the Sample 1 post-test all had very high scores. If the data were aggregated, item 12 was identified as problematic in all post-test samples.

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.* [12]. As such, IRT supports the identification of item 29 as problematic.

B. 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 [11]. 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, would not detect if items in an instrument were generally unfair. It can only detect when an item is functioning differently than the overall instrument.

1. Equal ability 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 1 shows this plot for the Sample 1 post-test. A line of slope 1 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. Figure 1 has three striking features: (i) most items are significantly unfair to women (the error bars do not overlap the fairness line); (ii) five items, 14, 21, 22, 23, and 27, stand out as substantially unfair to women by falling well off the fairness line; and (iii) 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. We focus on item difficulty for the remainder of the study.

To determine if the differences in performance in the CTT plot in Fig. 1 were statistically significant and to estimate effect sizes, the phi coefficient, ϕ, was calculated for each item and is included in Table III. The ϕ coefficient is equivalent to the two-point Pearson correlation coefficient for dichotomously scored items and provides a measure of effect size (Table I). The significance values for ϕ were calculated using the chi-squared test of independence on the two-by-two table of male and female correct and incorrect answers for each problem. The ϕ coefficient is related to χ^2 by $\phi = \sqrt{\chi^2/N}$ where N is the number of students. 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

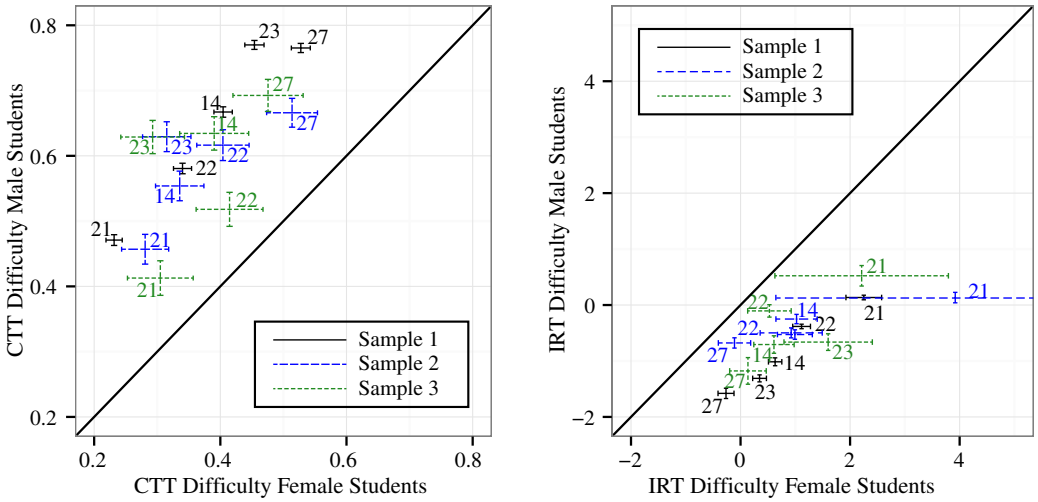

FIG. 3. CTT and IRT posttest difficulty scores for male and female students for problematic items from all samples. A line of slope one is drawn to allow comparison of male and female difficulty. The item number for each problem is also labeled. The IRT difficulty of Sample 2, item 23 is not labeled; the point overlays that of Sample 2, item 22.

items contains most of the items which will be identified as significantly unfair by DIF analysis.

A similar analysis was used to explore whether differences in the IRT difficulty coefficients were significant. The differences are characterized by Cohen's d (Table III). The results were similar to those using the CTT difficulty; the gender difference in items 14, 21, 22, 23, 26, and 27 was significant ($p < 0.001$) with a small to medium effect size. Table III also presents measures of the goodness of fit of the IRT model for men and women through the Cramer's V statistic. This analysis is described in the Supplemental Material [46].

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 refit removing this item. Parameter estimates changed very little; as such, the values for the original model including item 29 are reported.

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

The FCI pretest was analyzed using the same methods as the post-test; results are presented in the Supplemental Material [46].

2. *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. 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 III reports $\Delta\alpha_{\mathrm{MH}}$ for each item in Sample 1, stratified by total test score. 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 $\Delta\alpha_{\mathrm{MH}}$ and reported in Table III. The Lord's statistic results agreed with the high DIF classification provided by $\Delta\alpha_{\mathrm{MH}}$ 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.* [56]. 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.

C. Item-level analysis

The distribution of student answers for the five most unfair items of Sample 1 are shown in Table V. 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" [73]. Item 27 concerns a large box being pushed across a horizontal floor, and the preferred distractor across genders was that the box comes immediately to a stop.

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

		Response				
No.	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%

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). However, it is difficult to construct such an explanation that would not apply equally to items 9 and 15 (kicking a hockey puck and pushing a broken-down truck), which had a large DIF favoring women. 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 [74]. 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.

D. An unbiased Force Concept Inventory

To construct an unbiased version of the FCI, items were iteratively removed, $\Delta\alpha_{\mathrm{MH}}$ 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 $(73.1 \pm 19)\%$ and women scoring $(68.7 \pm 19)\%$. 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 (Table III). 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.

IV. DISCUSSION

This study sought to answer six research questions; these are addressed in the order proposed. We then consider larger patterns in prior research in light of our results.

A. 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? 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 (see Supplemental Material [46]) 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 Samples 1 and 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 IV, 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 nonproblematic 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 as both 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 III) 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. 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.* [56] (large DIF) and Osborn Popp *et al.* [47] (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, Fig. 4 of Ref. [56] 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 [57].

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 [2] 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. Some additional high DIF items were identified in reliability analysis but only after disaggregating by gender (see Supplemental Material [46]).

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 [57] 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 [7].

The reduced instrument did not significantly reduce the gender gap in Samples 2 and 3. An explanation may be found by comparing Fig. 1, Fig. 2, and the Sample 1 pretest plot (Fig. 1 in the Supplemental Material [46]). 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 postinstruction.

B. Insights into previous studies

Some studies have suggested that more interactive teaching methods lower the gender gap [34–36]; however, this effect has not been consistently reproduced [37–39]. Some research-based instructional methods were employed in the lecture portions of Samples 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 Samples 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 [34–39].

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 subsampling Sample 1 suggest 1000–1500 as a minimum sample size to clearly resolve gender disparities in FCI data sets 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 [3], 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.* [75] 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.

V. IMPLICATIONS

This work identified multiple questions within the FCI which were unfair to both 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 data sets 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. Unfortunately, item fairness has not received the same level of attention for the other commonly used mechanics conceptual inventory, the FMCE. If future research shows the FMCE does not contain a substantial collection of unfair items, it may be reasonable for institutions to use this instrument in the future.

VI. LIMITATIONS

While this research used data from four institutions combined to form three data sets, two of the data sets were too small to provide adequate statistical power to determine if some conclusions were general. The analysis should be conducted with additional large data sets to determine whether the conclusions are widely replicated.

Additionally, these results suffer from the same methodological constraints of all large-scale, quantitative studies where binary gender reporting is used. Coding all students (typically from institutional records) as male or female simplifies the complexity of gender identity, ignores the nuances of individual experiences, and (in the case of DIF) uses male students as the measure of "normal" against which female students are compared [8]. We chose to replicate these assumptions for the purpose of engaging with the long tradition of gender gap studies that follow this model. It is certainly not our intent to argue that quantitative analysis is the only or the best method for studying the gendered experiences of students in learning physics. However, ignoring even this "first order" model of gender can lead instructors to base conclusions about their students on flawed instruments.

VII. CONCLUSIONS AND FUTURE WORK

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 data sets 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 subscales should be sufficiently robust that the removal of some items leaves the validity and reliability of the instrument intact.

This work will be extended to the FMCE and the Conceptual Survey of Electricity and Magnetism (CSEM) [76] to determine how much, if any, of the gender gap reported in these instruments can be attributed to bias. This work should also be extended to investigate fairness for other underrepresented populations.

ACKNOWLEDGMENTS

This work was supported in part by the National Science Foundation as part of the evaluation of improved learning for the Physics Teacher Education Coalition, PHY-0108787. We appreciate the efforts of the instructors who contributed data to the three samples.

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

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

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

[4] J. Wang and L. Bao, Analyzing Force Concept Inventory with item response theory, Am. J. Phys. **78**, 1064 (2010).

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

[6] E. Brewe, J. Bruun, and I. G. Bearden, Using module analysis for multiple choice responses: A new method applied to Force Concept Inventory data, Phys. Rev. Phys. Educ. Res. **12**, 020131 (2016).

[7] A. Madsen, S. B. McKagan, and E. Sayre, Gender gap on concept inventories in physics: What is consistent, what is inconsistent, and what factors influence the gap?, Phys. Rev. ST Phys. Educ. Res. **9**, 020121 (2013).

[8] A. L. Traxler, X. C. Cid, J. Blue, and R. Barthelemy, Enriching gender in physics education research: A binary past and a complex future, Phys. Rev. Phys. Educ. Res. **12**, 020114 (2016).

[9] ETS Standards for Quality and Fairness, https://www.ets.org/s/about/pdf/standards.pdf, accessed 11/11/2017.

[10] M. Zieky, Fairness review in assessment, in *Handbook of Test Development*, edited by S. M. Downing and T. M. Haladyna (Lawrence Erlbaum, Hillsdale, NJ, 2006), pp. 359–376.

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

[12] N. Jorion, B. D. Gane, K. James, L. Schroeder, L. V. DiBello, and J. W. Pellegrino, An analytic framework for evaluating the validity of concept inventory claims, J. Eng. Educ. **104**, 454 (2015).

[13] 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).

[14] R. K. Thornton and D. R. Sokoloff, Assessing student learning of Newton's laws: The Force and Motion Conceptual Evaluation and the evaluation of active learning laboratory and lecture curricula, Am. J. Phys. **66**, 338 (1998).

[15] 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).

[16] 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).

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

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

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

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

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

[22] Y. Maeda and S. Y. Yoon, A meta-analysis on gender differences in mental rotation ability measured by the Purdue spatial visualization tests: Visualization of rotations (PSVT: R), Educ. Psychol. Rev. **25**, 69 (2013).

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

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

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

[26] N. M. Else-Quest, J. S. Hyde, and M. C. Linn, Cross-national patterns of gender differences in mathematics: A meta-analysis, Psychol. Bull. **136**, 103 (2010).

[27] X. Ma, A meta-analysis of the relationship between anxiety toward mathematics and achievement in mathematics, J. Res. Math. Educ. **30**, 520 (1999).

[28] J. V. Mallow and S. L. Greenburg, Science anxiety: Causes and remedies, J. Coll. Sci. Teach. **11**, 356 (1982).

[29] M. K. Udo, G. P. Ramsey, and J. V. Mallow, Science anxiety and gender in students taking general education science courses, J. Sci. Educ. Technol. **13**, 435 (2004).

[30] J. Mallow, H. Kastrup, F. B. Bryant, N. Hislop, R. Shefner, and M. Udo, Science anxiety, science attitudes, and gender: Interviews from a binational study, J. Sci. Educ. Technol. **19**, 356 (2010).

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

[32] E. Seymour, Undergraduate problems with teaching and advising in SME majors—explaining gender differences in attrition rates, J. Coll. Sci. Teach. **21**, 284 (1992).

[33] E. Seymour, The loss of women from science, mathematics, and engineering undergraduate majors: An explanatory account, Sci. Educ. **79**, 437 (1995).

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

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

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

[37] S. J. Pollock, N. D. Finkelstein, and L. E. Kost, Reducing the gender gap in the physics classroom: How sufficient is

interactive engagement?, Phys. Rev. ST Phys. Educ. Res. **3**, 010107 (2007).
[38] M. J. Cahill, K. M. Hynes, R. Trousil, L. A. Brooks, M. A. McDaniel, M. Repice, J. Zhao, and R. F. Frey, Multiyear, multi-instructor evaluation of a large-class interactive-engagement curriculum, Phys. Rev. ST Phys. Educ. Res. **10**, 020101 (2014).
[39] N. I. Karim, A. Maries, and C. Singh, Do evidence-based active-engagement courses reduce the gender gap in introductory physics?, Eur. J. Phys., DOI: 10.1088/1361-6404/aa9689 (2017).
[40] R. Henderson, G. Stewart, J. Stewart, L. Michaluk, and A. Traxler, Exploring the gender gap in the conceptual survey of electricity and magnetism, Phys. Rev. Phys. Educ. Res. **13**, 020114 (2017).
[41] S. L. Eddy and S. E. Brownell, Beneath the numbers: A review of gender disparities in undergraduate education across science, technology, engineering, and math disciplines, Phys. Rev. Phys. Educ. Res. **12**, 020106 (2016).
[42] L. Crocker and J. Algina, *Introduction to Classical and Modern Test Theory* (Holt, Rinehart and Winston, New York, 1986).
[43] N. Lasry, S. Rosenfield, H. Dedic, A. Dahan, and O. Reshef, The puzzling reliability of the Force Concept Inventory, Am. J. Phys. **79**, 909 (2011).
[44] C. Henderson, Common concerns about the Force Concept Inventory, Phys. Teach. **40**, 542 (2002).
[45] L. Ding and R. Beichner, Approaches to data analysis of multiple-choice questions, Phys. Rev. ST Phys. Educ. Res. **5**, 020103 (2009).
[46] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevPhysEducRes.14.010103 for further details on IRT and DIF models, pretest results, and reliability measures.
[47] S. Osborn Popp, D. Meltzer, and M. C. Megowan-Romanowicz, Is the Force Concept Inventory biased? Investigating differential item functioning on a test of conceptual learning in physics, in *2011 American Educational Research Association Conference* (American Education Research Association, Washington, DC, 2011).
[48] T. F. Scott and D. Schumayer, Students' proficiency scores within multitrait item response theory, Phys. Rev. ST Phys. Educ. Res. **11**, 020134 (2015).
[49] M. Planinic, L. Ivanjek, and A. Susac, Rasch model based analysis of the Force Concept Inventory, Phys. Rev. ST Phys. Educ. Res. **6**, 010103 (2010).
[50] J. C. Nunnally and I. H. Bernstein, *Psychometric Theory, Third Edition* (McGraw-Hill, New York, NY, 1994).
[51] R. K. Thornton, D. Kuhl, K. Cummings, and J. Marx, Comparing the Force and Motion Conceptual Evaluation and the Force Concept Inventory, Phys. Rev. ST Phys. Educ. Res. **5**, 010105 (2009).
[52] D. Hestenes and I. Halloun, Interpreting the force concept inventory: A response to March 1995 critique by Huffman and Heller, Phys. Teach. **33**, 502 (1995).
[53] P. Heller and D. Huffman, Interpreting the force concept inventory: A reply to Hestenes and Halloun, Phys. Teach. **33**, 503 (1995).
[54] P. W. Holland and D. T. Thayer, An alternate definition of the ETS delta scale of item difficulty, ETS Research Report Series **1985**, i (1985).
[55] P. W. Holland and D. T. Thayer, Differential item performance and the Mantel-Haenszel procedure, in *Test Validity*, edited by H. Wainer and H. I. Braun (Lawrence Erlbaum, Hillsdale, NJ, 1993), pp. 129–145.
[56] R. D. Dietz, R. H. Pearson, M. R. Semak, and C. W. Willis, Gender bias in the Force Concept Inventory? in AIP Conf. Proc. **1413**, 171 (2012).
[57] L. McCullough and D. E. Meltzer, Differences in male/female response patterns on alternative-format versions of FCI items, in *2001 Physics Education Research Conference Proceedings*, edited by K. Cummings, S. Franklin, and J. Marx (AIP Publishing, New York, 2001), pp. 103–106.
[58] L. McCullough, Gender, context, and physics assessment, J. Int. Womens St. **5**, 20 (2004).
[59] E. Mazur, *Peer Instruction: A User's Manual* (Prentice Hall, Upper Saddle River, NJ, 1997).
[60] Physport, https://www.physport.org. Accessed 8/8/2017.
[61] US News & World Report: Education Best Graduate Schools Physics, https://www.usnews.com/best-graduate-schools/top-science-schools/physics-rankings, accessed 4/30/2017.
[62] G. Novak, A. Gavrin, W. Christian, and E. Patterson, *Just-In-Time Teaching: Blending Active Learning with Web Technology*, 1st ed. (Addison-Wesley, Upper Saddle River, NJ, 1999).
[63] S. DeVore, J. Stewart, and G. Stewart, Examining the effects of testwiseness in conceptual physics evaluations, Phys. Rev. Phys. Educ. Res. **12**, 020138 (2016).
[64] P. M. Sadler, Psychometric models of student conceptions in science: Reconciling qualitative studies and distractor-driven assessment instruments, J. Res. Sci. Teach. **35**, 265 (1998).
[65] R. S. Lindell, *Enhancing College Students' Understanding of Lunar Phases, Unpublished doctoral dissertation* (University of Nebraska–Lincoln 2001).
[66] W. J. van der Linden, Unidimensional Logistic Response Models, in *Handbook of Item Response Theory* (CRC Press, Taylor & Francis Group, New York, 2016), Vol. 1, pp. 13–30.
[67] R. Zwick and K. Ercikan, Analysis of differential item functioning in the NAEP history assessment, J. Educ. Measure. **26**, 55 (1989).
[68] R. Zwick, *A Review of ETS Differential Item Functioning Assessment Procedures: Flagging Rules, Minimum Sample Size Requirements, and Criterion Refinement* (Educational Testing Service, Princeton, NJ, 2012).
[69] R Core Team, *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria 2017).
[70] D. Rizopoulos, ltm: An R package for latent variable modelling and item response theory analyses, J. Stat. Softw. **17**, 1 (2006).
[71] D. Magis, S. Béland, F. Tuerlinckx, and P. De Boeck, A general framework and an R package for the detection of dichotomous differential item functioning, Behav. Res. Meth. Instrum. Comput. **42**, 847 (2010).

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

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

[74] 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).

[75] 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. ST Phys. Educ. Res. **11**, 010112 (2015).

[76] D. P. Maloney, T. L. O'Kuma, C. Hieggelke, and A. Van Huevelen, Surveying students' conceptual knowledge of electricity and magnetism, Phys. Ed. Res., Am. J. Phys. **69**, S12 (2001).