

Is the Force Concept Inventory biased across the intersections of gender and race?

John B. Buncher; Jayson M. Nissen; Ben Van Dusen;
Robert M. Talbot

Accessibility Disclaimer:

For a more accessible version of this document, please submit an accessibility request form through the Montana State University Library website.

Is the Force Concept Inventory biased across the intersections of gender and race?

John B. Buncher¹, Jayson M. Nissen², Ben Van Dusen³, and Robert M. Talbot⁴

¹*Department of Physics, North Dakota State University, Fargo, North Dakota 58108, USA*

²*Department of Physics, Montana State University, Bozeman, Montana 59715, USA*

³*School of Education, Iowa State University, Ames, Iowa 50011, USA*

⁴*School of Education and Human Development, University of Colorado Denver, Denver, Colorado 80217, USA*



(Received 22 January 2024; accepted 26 February 2025; published 7 April 2025)

Research-based assessments (RBAs) allow researchers and practitioners to compare student performance across different contexts and institutions. In recent years, research attention has focused on the student populations these RBAs were initially developed with because much of that research was done with “samples of convenience” that were predominantly white men. Prior research has found that the Force Concept Inventory (FCI) behaved differently for men and women using differential item functioning (DIF) analysis. We extend this research in two ways. First, we test the FCI for DIF across the intersection of gender and race for Asian, Black, Hispanic, White, and White Hispanic men and women. Second, we apply the Eaton and Willoughby five-factor model of the FCI to interpret the results of the DIF analysis. We found large DIF on a large number of FCI items. The patterns of items with large DIF follows the five-factor model. The alignment of DIF with this factor structure, along with the measurement invariance of this factor structure across these ten social identities, indicates that the items on the FCI are likely not biased but are instead measuring real differences in physics knowledge among these groups. We frame these differences as educational debts that society owes to these marginalized groups that physics instruction needs to actively repay.

DOI: [10.1103/PhysRevPhysEducRes.21.010137](https://doi.org/10.1103/PhysRevPhysEducRes.21.010137)

I. INTRODUCTION

Research-based assessments (RBAs) have played a central role in measuring the impacts of physics courses and helped drive the adoption of research-based curricula [1,2]. Researchers often use RBAs to investigate the roles of racism, sexism, classism, and other forms of oppression in limiting science student success [3–10]. However, the RBAs researchers often use were primarily developed at research-intensive institutions [11] and often lack validation arguments that specifically examine their use with minoritized students [12–14], particularly from an intersectional perspective [15]. Intersectionality asserts that power relations of race, class, and gender, for example, are not discrete and mutually exclusive entities but build on each other and work together [16]. Without evidence that items and constructs on these RBAs act similarly across social identity groups, the claims about performance differences across social identity groups may be biased or inaccurate.

Traxler *et al.* [12] found that many items on the Force Concept Inventory (FCI) [17] performed differently for men and women and recommend removing 11 items from the 30-item instrument. An item performs consistently across two groups if members with the same ability from different groups have the same probability of answering an item correctly. Identifying members from two groups with the same ability poses a major challenge in these comparisons where systems of oppression have produced large differences between the groups being compared. We build on their work by examining the extent to which the items perform consistently across an intersectional set of ten social identity groups. We then examine whether items that perform inconsistently across groups are associated with five factors on the FCI identified by Eaton and Willoughby [18] in their EW5 model, which function equivalently across these ten social identity groups [19]. If the items function differently across groups and tend to occur on specific factors, then the inconsistent performance likely results from real differences in physics knowledge, not biased items, and instruction that addresses those factors could help to redress those inequities.

II. RESEARCH QUESTION

To better understand the potential for items on the FCI to bias student performance data across social identity

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

groups based on gender and race, we asked the following questions:

1. Which items on the FCI, if any, function differently across social identity groups?
2. To what extent do the items that function differently across social identity groups occur on distinct factors within the EW5 model?

III. DEFINITIONS

To support readers' interpretation of our research, Table I includes a selection of terms and definitions for statistical modeling and equity that we use.

IV. CONCEPTUAL FRAMEWORK

Intersectionality [23,25] posits that an individual's social identities (e.g., gender, race, and ethnicity) dynamically interact with the power structures in which they are embedded to create privilege and oppression [26]. The complex nature of the interactions between social identities and power structures invalidates the assumption that the impacts of racism, sexism, and other forms of oppression will have a simple additive nature [27] and highlights the importance of investigating outcomes across multiple social identity groups. For example, Black women often experience anti-Black racism in

different ways than Black men and sexism in ways that are different from White women [28–30]. In this investigation, intersectionality motivated us to disaggregate students' data across genders, races, and ethnicities to the extent the statistical methods allowed.

We use Ladson-Billings' [23,24] "educational debts" to conceptualize the differences in outcomes across social identity groups. We assume that groups have the same average potential across social identities and that differences in group outcomes are products of societal influences. Oppressive powers in our society cause inequities in group outcomes, rather than those outcomes being due to deficiencies in the individuals. These power structures exist within educational systems (e.g., school funding, access to advanced high school physics courses, and inequitable teaching practices) and broader societal contexts (e.g., inequitable wealth distribution, incarceration rates, and food access). By framing these inequities as society's educational debts, it places the onus on society to repay the debts, rather than on individuals for being "underprepared." As with monetary debt, unless society's educational debts are actively repaid then they continue to accrue interest and grow in size. Taking a perspective that ignores the role of race in education (i.e., a color-evasive perspective) tacitly accepts and thereby increases the debts that society owes to minoritized students.

TABLE I. Definition of some statistical and equity-related terms we use in the manuscript.

Term: Definition
<i>Research-based assessment (RBA)</i> : An assessment with reliability and validity arguments [2].
<i>Differential item function (DIF)</i> : A measure of the difference in group performance holding ability estimates constant. It is used to identify potential item bias.
<i>Artificial DIF</i> : When a single item exhibits real DIF, an equal but opposite amount of artificial DIF is distributed among the remaining items [20].
<i>Measurement invariance</i> : Measurement invariance testing assesses the psychometric equivalence of latent factors across groups or time [21].
<i>Classical test theory (CTT)</i> : A measurement model grounded in the assumption that each person has a true score that could be measured if there were no measurement error.
<i>CTT item difficulty</i> : The proportion of respondents correctly endorsing an item.
<i>Item response theory (IRT)</i> : A measurement model which describe the association between a latent trait and a respondent's probability of responding to an item in a particular way.
<i>Latent trait</i> : A variable that cannot be directly measured but can be inferred using a mathematical model of observable variables (e.g., Newtonian thinking or science identity) [22].
<i>IRT ability estimate</i> : The estimate of a respondent's latent trait, e.g., Newtonian thinking, as measured by a test.
<i>IRT item difficulty</i> : The ability level where a respondent has a 50% chance of correctly endorsing an item.
<i>Intersectionality</i> : Intersectionality asserts that the power relations of race, class, and gender, for example, are not discrete and mutually exclusive entities but rather build on each other and work together [16].
<i>Social identity group</i> : A group defined by physical, social, and mental characteristics of individuals. For example, race or ethnicity, gender, social class or socioeconomic status, sexual orientation, (dis)abilities, and religious beliefs.
<i>Educational debt</i> : "Education debt is the foregone schooling resources that we could have (should have) been investing in (primarily) low-income kids, which deficit leads to a variety of social problems (e.g., crime, low productivity, low wages, low labor force participation) that require ongoing public investment." [23] This reframes inequities in group outcomes from deficits in student abilities to debts that society owes to minoritized students due to racism, sexism, and classism [23,24].

A. Positionality

The members of our team each brought unique lived experiences and perspectives to this research. We provide positionality statements for each author to contextualize the perspectives represented in this research.

John Buncher: I identify as a continuing-generation White cisgender man. I earned a bachelor's, master's, and Ph.D. in physics, have engaged in physics research, and have been a physics faculty member for the entirety of my professional career. I grew up in a middle-income household. Experiences in graduate school and my nearly all-male physics course in high school motivated me to reflect on and help dismantle the ongoing structural oppression present in physics and higher education at large.

Jayson Nissen: Identifying as a first-generation, White, cisgendered, man provides me with opportunities denied to others in American society. I have a Ph.D. in physics, teach physics, and have worked as a chemist, which focuses my work on the sciences. Because I focus on quantitative research, I want to make sure the tools and methods informing my research are valid for the purposes I use them for.

Ben Van Dusen: I identify as a continuing-generation White cisgender man. I earned a bachelor's degree in physics, have engaged in physics research, was a high school physics teacher, and now have a Ph.D. in education and prepare future science teachers. I was raised in low-income households but now earn an upper-middle-class income. People with privileges similar to mine have created and maintained our society's unjust power structures. As a person with privilege, I believe it is my obligation to use it to dismantle oppressive systems. I also recognize that my privilege, however, limits my perspective on the lived experiences of minoritized individuals.

Robert (Bud) Talbot: I identify as a continuing-generation, White, cisgender man. I was raised in a middle-income Midwestern household with working parents who were first and second generation out of

farming. I have earned bachelors, masters, and a Ph.D. in science education from flagship public research-intensive institutions. I taught middle and high school physics, earth science, and general science in both rural and urban schools before earning my Ph.D. In my work as an educator and researcher, I strive to use my privilege in the service of elevating the voices of others who are underrepresented or minoritized by our society.

V. BACKGROUND

A. Instrument validation

Developing a validity argument for an instrument involves collecting and analyzing evidence related to multiple aspects of validity [31]. Although standards for instrument validity are evolving [32], instrument bias has always been a central aspect of validity. At a minimum, bias in an instrument (e.g., toward a specific subpopulation or group) can present a source of construct-irrelevant variance that can exaggerate or obscure differences between groups, which can lead to wasted resources or the perpetuation of injustices. Fairness, as it relates to bias, involves ensuring that respondents have the opportunity to demonstrate their knowledge, and that the scores resulting from the test will be valid for the intended uses [31].

B. Differential item function analysis

Differential item functioning (DIF) is a general term for describing how test items may perform differently or unexpectedly for two groups [33,34]. DIF analysis refers to the two groups as the focal and reference group. In our work, those groups are defined by social identifiers for gender and race or ethnicity. By comparing item scores for individuals having the same overall ability across social identifier subgroups, we can see if the items are functioning differentially for certain groups.

Items functioning differentially for groups, as indicated by DIF, is a strong indicator of potential bias. DIF, however, is neither necessary nor sufficient for an item or a test to be biased or unfair. Real and meaningful differences can exist between groups that create DIF. Holland [35] addresses the role that researchers can play in creating tests that minimize or maximize differences between groups, and thereby skew DIF scores on items, in discussing the "Golden Rule law suit settlement" [35,36]. The Golden Rule procedure attempted to minimize the score differences between Black and White test takers by choosing only those test questions that minimized the empirical performance difference between these two groups. Holland [35] applied this rule to show how to construct tests that maximize or minimize the differences between two groups.

In so doing, Holland demonstrates that while DIF is useful for identifying problematic items, DIF is neither necessary nor sufficient for determining if a test or test items are fair. DIF may identify unfair items and DIF can also identify real differences. Unfair, biased tests can have minimal DIF so long as all of the items are similarly unfair toward one group. For example, Holland [35] designed a test that maximized the difference in scores between two groups by creating a short assessment made of only the items from a longer assessment that favored a particular group.

Researchers use many statistical methods to examine DIF. Under classical test theory measurement models (CTT), researchers often use the Mantel-Haenszel (MH) method [37]. This method uses a student's overall assessment score as an estimate of their ability and compares how each group performs on an item at each ability level.

Item response theory (IRT) measurement models estimate item properties (e.g., difficulty, discrimination, guessing) and a person's latent trait of interest (e.g., Newtonian thinking) from their responses to a set of items. These estimates are probabilistic in nature, and the person's latent trait of interest is often referred to as a person's ability estimate. Rasch models are a class of IRT models that estimate one item property, the item difficulty, in addition to the person's latent trait of interest.

Rasch models provide an affordance over CTT models of person- and item-invariance across testing populations. In plain words, person and item estimates should not vary across different populations within a latent trait-based Rasch model, provided that the model fits the data. If, however, modeled item difficulty estimates vary across samples or groups of students by much more or less than the amount observed in the data, then that is evidence of DIF within a group or population. These differences between modeled and observed item difficulty estimates manifest themselves in the item fit statistics, specifically infit and outfit. Variation in latent ability within groups (manifested in test scores) greatly impacts any investigation of item bias under a classical test theory (CTT) modeling framework unless those abilities within the group are representative of the universe of possible abilities. Ability estimates derived from a Rasch model are independent of the range of observations drawn from the sample. Therefore Rasch-based DIF analyses are stronger than CTT-based item bias investigations of test validity where issues of fairness and equity are of primary concern [20,38].

One limitation of DIF analysis is the presence of artificial DIF. As demonstrated by Andrich and Hagquist [20], when a single item exhibits real DIF in both the Mantel-Haenszel and Rasch models, an equal but opposite amount of artificial DIF is distributed among the remaining items. A large number of items exhibiting large amounts of DIF can cause enough artificial DIF to accumulate on otherwise DIF-free items such that they are flagged as having DIF in the opposite direction of the items with real DIF. The presence of artificial DIF is a reason for researchers to exercise caution in how DIF analyses are interpreted.

Central to a Rasch-based analysis of instrument bias or DIF is the notion of *anchoring*. Anchoring is the process of fixing the estimated parameters for items identified as unbiased so that the parameters for other items can be freely estimated and evaluated for DIF relative to the anchored items, which we detail in the Methods.

C. Existing DIF studies on physics concept inventories

Previous research using DIF to investigate potential item bias on physics concept inventories has focused on gender differences among respondents [12,13,39,40]. In those works, research questions were posed about whether properties of the test itself influenced performance by gender.

Osborn Popp *et al.* [39] used Rasch modeling of the FCI to look for differences between the estimated item difficulty parameters between gender groups and used this as evidence of DIF. They assessed these differences using *t* tests. Using this method, they found “no clear trend favoring males” but did identify three items with “substantial” DIF, two of which favored males and one of which favored females. The two items favoring males (23 and 14) had been shown in other research [12,41] as having large gender differences in performance.

Using both a CTT framework and the MH statistic, as well as an IRT framework using Lord's statistic, Traxler *et al.* [12] investigated DIF and gender fairness on the FCI. Notably, these researchers did not use a Rasch IRT model but rather a two parameter logistic model (2PL). Using a multiparameter model allows for more relaxed fitting of the data to the model, as some of the observed variance is partitioned into a second item parameter (discrimination). Also, the additional item parameter estimates in multiparameter models do not exist on the same scale as the person parameter estimates, as they do under the Rasch model. In their CTT-based analysis (using MH), Traxler *et al.* [12] found eight items on the FCI that exhibited large amounts of DIF with only two of those favoring women. They also identified seven additional items that showed small to moderate amounts of DIF. In their IRT-based analysis (using Lord's statistic), the researchers also identified the same eight items as having large DIF (and additionally, item 29). The items identified as having small to moderate DIF were less consistent between the two analyses.

Traxler *et al.* [12] recommended removing 11 items from the FCI based on these results. They compared gender differences for three samples of students when they removed ten items because of the item's DIF. In one sample, removing the ten items halved the gender difference. In the other two samples, it had no effect.

Henderson *et al.* [13] examined item-level gender differences on the FMCE and CSEM using CTT item difficulty and discrimination. They investigated evidence of DIF using the MH statistic. Although they did find

evidence of DIF favoring males on the FMCE, “The number of items that demonstrated large DIF in the FMCE was much smaller than the eight large DIF items initially identified by Traxler *et al.* [12] in the FCI.” They conclude that the FMCE is much less gender unfair than the FCI. In a follow-up study, Henderson *et al.* [40] used hierarchical linear modeling to investigate the relationship between various factors and the observed differences between male and female scores on these concept inventories. They found that “correcting the instruments for fairness explained differing amounts of the gender differences.” In one of their FCI samples, fairness accounted for 30% of the gender differences. This was consistent with their previous DIF analyses.

D. The Force Concept Inventory

Hestenes *et al.* [17] developed the FCI to investigate student understanding of Newtonian physics. The FCI consists of 30 multiple choice questions and is available from PhysPort [42] or on the LASSO platform [43]. Researchers have used the FCI to investigate the effectiveness of instructional techniques [44–48] and equity in courses [7,49,50].

Hestenes *et al.* [17] proposed that the FCI covered six areas of Newtonian forces. Findings from multiple studies, however, suggest that it examines five factors [14,19,51,52]. Eaton and Willoughby [51] found that their five-factor model (EW5), shown in Table II, provided the best fit for the data. Morley *et al.* [19] found that this model fit the FCI well and provided equivalent measurement across the ten social identity groups in this study.

In addition to the work by Traxler *et al.* [12] that identified items with DIF, several studies have identified items on the FCI with poor psychometric properties. Wang and Bao [53] applied a three-parameter item response model and Planinic *et al.* [54] applied a Rasch model. As Morley *et al.* [19] discuss, applying the EW5 model to their results indicates that factors 1 and 2 tend to have easier questions with lower discrimination and higher likelihoods of guessing the right answer than the other three factors. Both Planinic *et al.* [54] and Morley *et al.* [19] found that the FCI may function differently for populations with lower test scores.

TABLE II. Eaton and Willoughby [51] five-factor (EW5) model.

Factor	Description	Items
F1	Newton’s first law & kinematics	6, 7, 8, 10, 20, 23, 24
F2	Newton’s second law & kinematics	9, 12, 14, 19, 21, 22, 27
F3	Newton’s third law	4, 15, 16, 28
F4	Identification of forces	5, 11, 13, 18, 30
F5	Superposition	17, 25, 26

VI. METHODS

A. Data collection

We collected the data from the Learning About STEM Student Outcomes (LASSO) platform’s anonymous research database [43,55]. Morley *et al.* [19] also used an overlapping dataset with the one used in this study in their investigation of the measurement invariance of the Eaton and Willoughby five-factor model across the same ten social identity groups investigated in this study. LASSO is an online assessment platform that administers low-stakes RBAs and provides instructors with analyses of student performance. Our analysis included student FCI scores from 354 calculus-based first-semester college physics courses across 45 institutions. The institutions were distributed across the four census tracts in the Northeast (13), South (12), Midwest (7), and West (13). We used the Carnegie classification of institutions of higher education (CCIHE) public 2021 database to characterize the 44 institutions in the database. The one institution not in the CCIHE was a branch campus that was physically separated from its main campus and that likely acted as a different type of institution than its branch campus, i.e., a regional primarily undergraduate institution as opposed to a national research-intensive institution. The dataset included data from 19 high or very high research-intensive institutions, 9 minority-serving institutions that were all Hispanic-serving, and 5 primarily associate degree granting institutions. Thirty were public and 14 were private not for profit institutions. The data included 21 large, 18 medium, and 5 small institutions.

We took several steps in preparing the data. We only included students who were enrolled in a calculus-based introductory physics course. To ensure the data were representative of students making a genuine attempt at the FCI, we only included students who took more than 5 min to complete the exam and completed more than 80% of the questions on the exam. The data were further filtered to remove social identities with too few students for the analyses by retaining students who did not select both man and woman for gender, students who selected (for example) man and transgender were included as men, and selected only one ethnicity or race, except for the case of White and Hispanic, which we report as White Hispanic. Table III includes sample size and descriptive statistics for each of the ten social identity groups.

B. Data analysis

The analysis used both IRT and CTT, via the Mantel-Haenszel method, to test for DIF. Both analyses compare the relative difficulties of items for a minoritized group (the focal group) in physics to White men (the reference group). The definition of item difficulty is detailed for each method and in Table I. We used both IRT and CTT approaches to mirror the work by Traxler *et al.* [12] and to

TABLE III. Descriptive statistics for the pre- and post-tests. The pretest sample is for any students who completed a pretest. The post-test is for any student who completed a post-test. Both pretests and post-tests were subjected to the filtering rules outlined in the text.

Gender	Race	Pre			Post		
		Mean	SD	<i>N</i>	Mean	SD	<i>N</i>
Men	Asian	45.1	23.5	722	59.1	26.0	633
Men	Black	35.1	19.2	269	50.4	23.4	218
Men	Hispanic	35.7	17.2	363	54.5	21.7	279
Men	White	51.9	22.0	4438	68.2	22.7	3651
Men	White Hispanic	40.9	20.2	508	59.9	24.0	440
Women	Asian	37.5	22.1	362	57.6	25.6	302
Women	Black	30.3	20.5	213	50.9	22.3	184
Women	Hispanic	29.0	19.4	127	51.4	26.6	106
Women	White	38.1	20.1	1465	59.9	23.6	1202
Women	White Hispanic	31.9	16.9	205	53.0	23.6	183

TABLE IV. Contingency table for subgroup *k* in the Mantel-Haenszel test.

	Correct	Incorrect	Total
Reference	A_k	B_k	$A_k + B_k$
Focal	C_k	D_k	$C_k + D_k$
Total	$A_k + C_k$	$B_k + D_k$	$A_k + B_k + C_k + D_k$

provide a contrast for the approach we took with anchoring in the IRT analysis.

1. Classical test theory (CTT)

The Mantel-Haenszel method measures DIF using 2×2 contingency tables to compare item scores between groups for all items on a test [37]. The range of total scores on the test is divided into *k* intervals that then serve as the basis for matching members of the groups. Contingency tables for each interval are then constructed to compare groups on item performance, as shown in Table IV.

Typically, the reference group is the majority group and the focal group is the minority group. In this work, White men are our reference group, and each of the nine other social identity groups was a focal group. Using White men as the reference group, especially without explicitly stating so, can reify dominant groups as representing the normal state [56]. DIF analysis seeks to break that paradigm by naming the minoritized group as the focal group and explicitly addressing potential bias and unfairness in assessments.

Each contingency table is used to compute an odds ratio α_k for that subgroup on a given item, given by

$$\alpha_k = \frac{A_k/(A_k + B_k)}{B_k/(A_k + B_k)} \cdot \frac{C_k/(C_k + D_k)}{D_k/(C_k + D_k)} = \frac{A_k D_k}{B_k C_k}. \quad (1)$$

A weighted sum is performed over all subgroups to find the common odds ratio α , given by

$$\alpha = \sum_k \alpha_k w_k / \sum_k w_k, \quad (2)$$

where the weight is given by $w_k = B_k C_k / (A_k + B_k + C_k + D_k)$. If no DIF is present, then $\alpha = 1$. Since odds ratios can be difficult to interpret and range from 0 to infinity, α is often log-transformed to be symmetric about 0 via

$$\Delta\alpha_{MH} = -2.35 \ln \alpha, \quad (3)$$

with positive values indicating items that favor the focal group and negative values indicating items that favor the reference group. Here we match the notation used by Traxler *et al.* [12], but we note that elsewhere $\Delta\alpha_{MH}$ is referred to as “MH D-DIF” [57].

A widely used scale for $\Delta\alpha_{MH}$ is the one adopted by the Educational Testing Service [57,58]. Items are classified as A, B, or C. An A item is one where the Mantel-Haenszel chi-square statistic is not significant at the 0.05 level or if $|\Delta\alpha_{MH}| < 1$. A C item is one in which the $|\Delta\alpha_{MH}|$ statistic is significantly greater than 1 at the 0.05 level and $|\Delta\alpha_{MH}| \geq 1.5$. A B item is one that is neither classified as an A item nor a C item. In this work, we use these established cutoffs for $\Delta\alpha_{MH}$ without the *p*-value restrictions and classify items with $|\Delta\alpha_{MH}| \geq 1.5$ as having large DIF, and items with $1.0 \leq |\Delta\alpha_{MH}| < 1.5$ as having moderate DIF. All Mantel-Haenszel tests were performed with version 5.1 of the difR library in the R software package [59].

2. Item response theory (IRT)

In our IRT analysis, we used the Rasch model (sometimes referred to as a one-parameter logistic model or 1PL model). The Rasch model allowed us to model groups with smaller sample sizes than the 2PL model [60], which aligned with our intersectional framework. Rasch analysis models the probability that participant *i*, with ability estimate θ_i , answers item *j* correctly as

$$P_{ij}(\theta_i) = \frac{1}{1 + e^{-(\theta_i - b_j)}}. \quad (4)$$

Here, b_j is the difficulty of item *j* and is the ability estimate level needed by a respondent to have a 50% chance of answering the item correctly. For a given comparison between two groups, the parameters b_j for each item (per group) and the ability θ_i (per student) are estimated using maximum likelihood methods.

Comparing the difficulty parameter across groups requires establishing a reference scale. This is done by identifying *anchor items* and constraining the difficulty

parameter of those items to be the same for the two groups. These anchors allow estimating the other item parameters and the ability scores on this common scale. Careful consideration of the anchor items is critical for DIF analysis. If items in the anchor contain DIF (referred to as contamination), the common scale is effectively shifted, possibly resulting in DIF-free items identified as containing DIF, while items with DIF go undetected. A shorter list of items, typically four, can reduce the risk of contamination, but this can limit the statistical power to detect DIF [61].

Methods of selecting the anchor items vary, with the optimal method depending on the psychometric model used (Rasch, 2PL, 3PL, etc.), the length of the assessment, the scoring of the assessment (polytomous vs dichotomous), the amount of DIF in items, the distribution of the direction of DIF in items, and the percentage of items exhibiting DIF [61]. Many of these methods rest on the implicit assumption that the majority of items are DIF free, and simulations that test the robustness of these methods when the number of DIF items is 60% or higher are rare [62]. While many recommendations in the literature exist, the choice of anchoring method often comes down to choices made by the researcher. For example, Strobl *et al.* [62] propose using the Gini Index from economics to select anchor items in a procedure that shifts the item difficulties to maximize the amount of DIF in the fewest number of items. Kopf *et al.* [61] advocate for the mean p -value threshold (MPT) method, which iteratively adds anchor items whose p value on a preliminary statistical test for DIF is below some threshold. This method rests on the assumption that the majority of items on the exam perform similarly across the two groups and only a few items have meaningful DIF.

For the FCI, extensive research shows that overall performance before and after instruction varies across race and ethnicity [3,7] and gender [3,7,41,63–68], with differences as large as 20 percentage points or one standard deviation, which we conceptualized as societal educational debts. To investigate the FCI with IRT DIF analyses, we needed to determine how much of the difference between groups is attributable to the educational debts society owes to the minoritized groups while accounting for biased items exaggerating the size of these educational debts.

To account for the educational debts, we needed a method to quantify the amount of educational debt each group would have in the specific domain of Newtonian physics as measured by the FCI. We chose to use the group average of the person latent trait estimates from an initial Rasch model which assumed that the FCI performs identically across all groups, i.e., no DIF is present. This approach puts the measured educational debts on an identical scale as the item difficulties, which aids in anchor item selection. A disadvantage is that any items with significant DIF not due to educational debt may affect the estimate of educational debts. This drawback occurs for

any DIF analysis that assumes the overall test functions similarly across groups. We addressed this possibility with two analyses of the post-test educational debts: one included all of the questions on the FCI and the second used a 19-item FCI that excluded the items identified as problematic by Traxler *et al.* [12].

To measure the educational debts specific to Newtonian thinking as measured by the FCI post-test, we ran a Rasch model treating two groups (i.e., White men and Black women) as a single group (i.e., each item j has a single difficulty parameter b_j). The latent traits (ability estimates) θ for each student are then extracted from the model. The difference in the average latent trait, $\bar{\theta}$, between the two groups is a measure of the educational debt, ED :

$$ED = \bar{\theta}_{\text{minoritizedgroup}} - \bar{\theta}_{\text{Whitemen}}. \quad (5)$$

The results are listed in Table V and show that the educational debts determined by the 30-item FCI are within 0.11 standard deviations of the educational debts determined by the 19-item FCI across all 9 groups. Because of these small differences, we focused on the educational debts as measured by all of the questions on the FCI post-test.

To identify anchor items, we then ran the Rasch model on the post-test responses and estimated item difficulties for each item separately for each group. This is equivalent to assuming that the distribution of latent traits in each group has the same average and standard deviation, but that the items could have different difficulties across groups. We then calculated the difference in item difficulties, and chose as anchors the three items that had a difference in difficulty between White men and each focal group that were closest to the educational debt measured for that

TABLE V. Educational debts across race or ethnicity and gender, as compared to White men, using the difference in the average latent trait as measured on the post-test. Uncertainties are 1 SE. Negative values indicate a debt owed to the respective group. The 30 items column uses all FCI items to measure the educational debt, while the 19 items column excludes items 6, 9, 12, 14, 15, 21, 22, 23, 24, 27, and 29 which were found to be potentially problematic by Traxler *et al.* [12].

Race	Gender	Educational debt			
		(30 items)		(19 items)	
White	Women	−0.46	0.04	−0.39	0.05
Asian	Women	−0.56	0.09	−0.45	0.10
Black	Women	−0.98	0.10	−0.92	0.10
Hispanic	Women	−0.87	0.16	−0.84	0.18
White Hispanic	Women	−0.84	0.10	−0.77	0.11
Asian	Men	−0.46	0.07	−0.44	0.07
Black	Men	−0.98	0.09	−1.00	0.11
Hispanic	Men	−0.78	0.08	−0.80	0.08
White Hispanic	Men	−0.43	0.07	−0.47	0.08

group. This choice assumes that items with a difference in difficulty close to the measured educational debt provide equivalent measures of student performance across both groups. We excluded item 29 as a possible anchor due to its poor fit in prior studies [12].

3. Rasch model assumptions

The Rasch model assumptions include unidimensionality of the measured trait, monotonicity, local independence, and item invariance [60]. The last, item invariance, assumes items operate the same across different groups of students and is the focus of our research questions.

Evidence for the unidimensionality in our work comes from two sources. First, almost all of the infit and outfit statistics fell between 0.5 and 1.5 and thus were acceptable [69], see Table XV. The only exceptions are for item 29, which Traxler *et al.* point out is a problematic item. Second, we conducted a confirmatory factor analysis using lavaan [70]. For a unidimensional model including all of the data, the robust comparative fit index (CFI) = 0.97, robust Tucker-Lewis index (TLI) = 0.97, and robust root mean square error of approximation (RMSEA) = 0.04, all of which are acceptable.

Monotonicity refers to the likelihood that a student's correct answer to a question increases as the student's ability increases. We generated item response curves for each item and found acceptable monotonicity.

Local item independence assumes that if we control for ability, the performance on each item is not conditionally dependent on performance on another item. We tested local independence using the mirt package to generate LD and Cramér's V statistics. Many of the LD statistics exceeded the thresholds suggested by Chen and Thissen [71]. This aligned with the findings of Eaton *et al.* [72] that the FCI lacked local independence but conflicts with the arguments of Wang and Bao [53]. The impact of exceeding these LD statistics on the FCI, however, is not clear. Cramér's V provides an effect size to help interpret the potential effects of violating this assumption. Cohen [73] provides cutoffs of <0.1 as negligible and <0.3 as small. Of the 435 Cramér's Vs, none exceeded the 0.3 threshold for a small effect and 28 were small and positive and 35 were small and negative. Of these 35 small negative values, 33 included item 15 or 29. Traxler *et al.* [12] make no mention of local independence, but the negative values for items 15 and 29 may align with their findings that these items favored women and should be removed from the exam. The items they found favoring men, however, did not have small positive Cramér's V values in our analysis. While this lack of complete local independence may spread out the abilities that the Rasch model outputs, the effect is likely small, likely does not impact our analyses, and is likely similar to other studies using the FCI.

4. IRT-DIF effect size

To quantify the amount of DIF present on a given item across two groups, we used the expected score standardized difference (ESSD) [74]. We first define the expected score for item i , which is the sum of probabilities of each of the m possible responses times the value of that response:

$$ES_{s(\hat{\theta})i} = \sum_{k=1}^m P_{ik}(\hat{\theta})X_{ik}. \quad (6)$$

Here, $ES_{s(\hat{\theta})i}$ is the expected score of the s th student with an estimated ability level $\hat{\theta}$, and $P_{ik}(\hat{\theta})$ is the probability of a student with an estimated ability level $\hat{\theta}$ to choose option k .

When fitting the IRT model to two different groups, each group will end up with their own set of item parameters γ (e.g., item difficulties), one for the focal (minority) group γ_F and one for the reference (majority) group γ_R . For the i th item, the ESSD is calculated as the difference between the expected score for the focal group's estimated ability levels using their own item parameters and the expected score for the focal group's estimated ability levels using the item parameters of the reference group, in units of the pooled standard deviation:

$$ESSD_i = \frac{\overline{ES}_{(\gamma_F)} - \overline{ES}_{(\gamma_R)}}{SD_{\text{ItemPooled}}} \quad (7)$$

with the item-level pooled standard deviation given by

$$SD_{\text{ItemPooled}} = \sqrt{\frac{(N_F - 1)SD_{ES(i|\gamma_F)} + (N_R - 1)SD_{ES(i|\gamma_R)}}{2(N_F + N_R - 2)}},$$

where the size of the focal group is given by N_F . The ESSD is expressed in units of the standard deviation, making it analogous to Cohen's d . This standardization of the ESSD allows us to compare it to other standardized values, such as student ability scores and the educational debts that we measured as the differences in these ability scores across social identity groups. Uncertainties for the ESSD values were computed via bootstrapping 1000 trials. The idea behind the ESSD is not to calculate the difference in expected scores across the two groups, rather it is to calculate the standardized difference in expected score on a single item for the *focal* group using the item parameters of *both* groups.

For the ESSD, since it is an effect size in units of the pooled standard deviation similar to Cohen's d , we used the common medium and large cutoffs of 0.5 and 0.8. All IRT calculations were done using version 1.38 of the mirt library in the R software package [75].

5. Reporting results

For comparison to Traxler *et al.* [12], Tables VI–VIII include indications of statistical significance ($p < 0.05$) calculated using a Bonferroni correction [76]. We report on DIF, however, without regard to the statistical significance in alignment with the American Statistical Association’s statement on p values [77], “Scientific conclusions and business or policy decisions should not be based only on whether a p value passes a specific threshold.” Basing decisions on a strict cutoff for p values results in a biased literature base [78]. Instead, we balance the size of the DIF, the consistency across comparisons, and the uncertainty in the DIF [78–80]. This decision allows the results to show items with moderate or large DIF even if the results have large uncertainties due to the small sample sizes in our intersectional groups.

C. Limitations

A central limitation of DIF analysis is that it is not able to directly identify biased items. DIF analysis identifies items that groups appear to over- or underperform on relative to the other items and other groups. These differences can arise from one of two sources. If the differences in item difficulties arise from differences in student knowledge about concepts not related to the construct of interest, then the question is not measuring what it is intended to measure and has bias. If the differences in item difficulties arise from true differences in student knowledge about concepts related to the construct of interest, then there is no bias. So while DIF analysis flags items as potentially biased, further analysis of the item wording and student interpretations of them are required to determine if the items are truly biased.

To avoid generating uncertain and potentially misleading findings, we used a minimum subgroup sample size of 100 people. Therefore, groups that did not have at least 100 people in our sample were not included. Some groups that we had to exclude include: nonbinary gender, Pacific Islander, Middle East/North African, and most multiracial groups. The only multiracial group that had a large enough

population in our data to be included was White-Hispanic students. Thus, our findings do not speak to potential bias in items for the groups not included in our study.

While the data in this study comes from a range of institutions, there are still limits to its generalizability. The LASSO platform’s database has more diversity in its institutional contexts than the literature [48], but it still skews toward research-intensive institutions. Our work would be strengthened by additional examinations of student answers specifically at minority-serving institutions, Hispanic-serving institutions, tribal colleges and universities, and historically Black colleges and universities.

VII. FINDINGS

A. CTT based DIF

Table VI and Fig. 1 present the results of the Mantel-Haenszel DIF tests. Of the 30 items on the FCI, 26 had medium-to-large DIF in at least one of the 9 comparisons, either favoring White men or the minoritized group. Of these, 19 items were statistically significant in at least one comparison, but we report all items that had medium-to-large effect sizes in Table VI to examine trends across groups and for transparency [77].

The size and frequency of items exhibiting DIF were greater when comparing White men to women than to any other group of men. This is reflected in the far fewer items in the last four rows of Table VI and the much larger spread in the $\Delta\alpha_{MH}$ values for women than for men in Table IX in the Appendix. Items 12, 14, and 23 consistently favored White men over all groups of women, with items 6, 10, and 27 favoring White men over four out of five groups of women. As shown in Fig. 1, these six items came from the first or second law factors of the EW5 model. Very few items consistently favored the minoritized groups over White men. Item 29 was one (favoring all groups except Asian men and White Hispanic men over White men), and item 9 favored women (excepting White non-Hispanic women) over White men.

TABLE VI. Item DIF across race or ethnicity and gender using classical test theory. Items with $1 < |\Delta\alpha_{MH}| < 1.5$ are classified as having a medium DIF, items with $|\Delta\alpha_{MH}| > 1.5$ are classified as having large DIF. Asterisk (*) indicates $p < 0.05$ (Bonferroni-corrected).

Focal group		DIF favoring focal group		DIF favoring White Men	
Race	Gender	Large	Medium	Medium	Large
White	Women		29	6 , 10 , 12	14 , 23
Asian	Women	17 , 18 , 25	4 9 , 15 , 20 , 26 , 28 , 29	6 , 12	14 , 23 , 27
Black	Women	2 , 4 , 9 , 15 , 17 26	29	7 , 10 , 12 , 24	6 , 8 , 14 , 23 , 27
Hispanic	Women	4 , 15 , 17 26 , 29	5 , 9 , 16	8 , 10 , 12	14 , 23 , 27
White Hispanic	Women	29	4 , 9 , 17 , 25	3 , 13 , 22	6 , 10 , 12 , 14 , 23 , 24 , 27
Asian	Men		17 , 18 , 26	27	
Black	Men	17 , 28 , 29	2 , 30	14	
Hispanic	Men	4 , 17 , 29	28		
White Hispanic	Men			12	

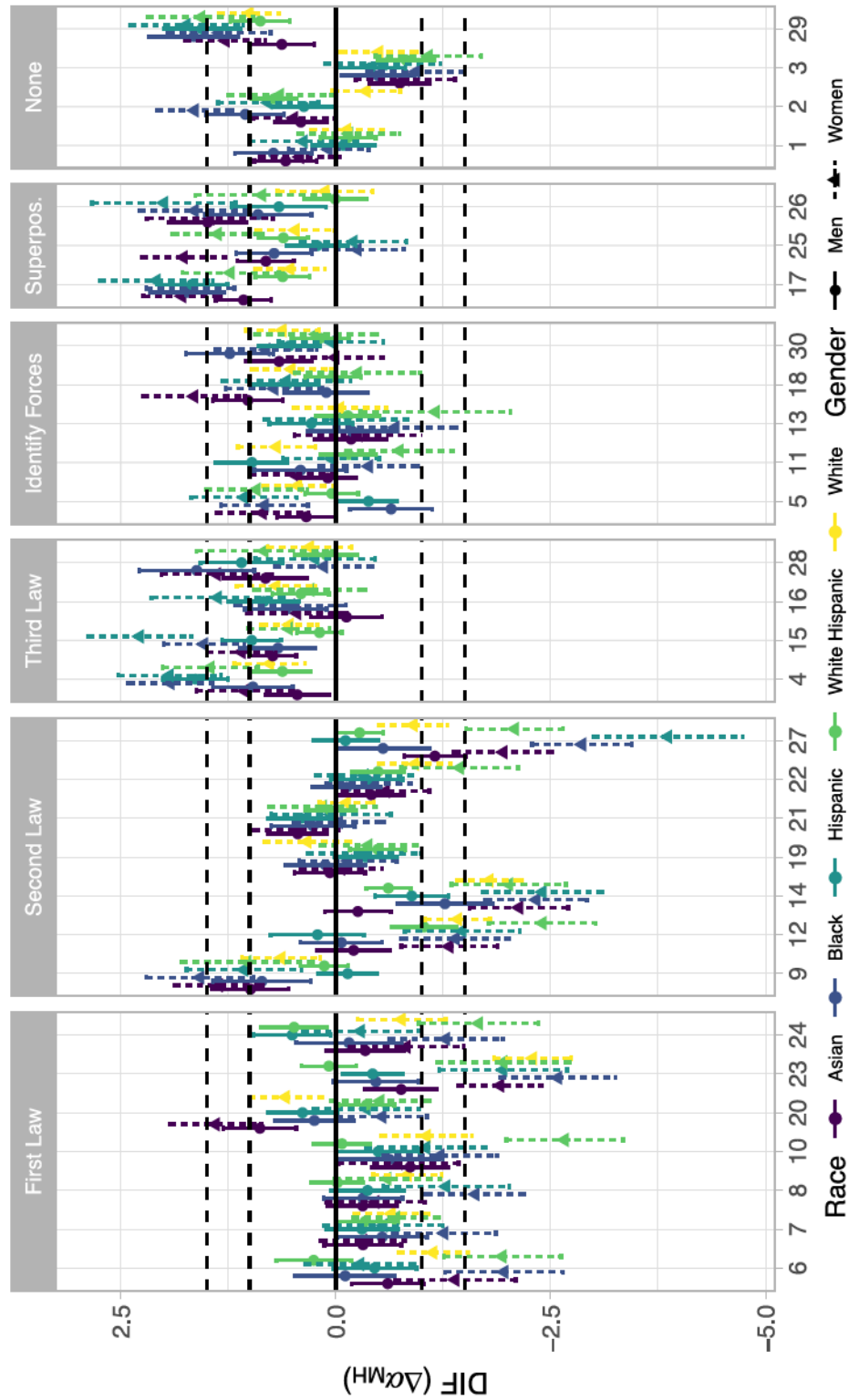


FIG. 1. $\Delta\alpha_{MH}$ values of the 30 FCI items organized by the factors of the EW5 model [18]. Error bars represent 1 standard error. The standard cutoffs of 1.0 and 1.5 for medium and large values of $|\Delta\alpha_{MH}|$ are shown with dashed lines. Items with $\Delta\alpha_{MH} < 0$ have DIF in favor of White men.

TABLE VII. Item DIF across race or ethnicity and gender using Item Response Theory. Educational debts were accounted for using three anchors. Cutoffs for medium and large DIF were ESSD scores of 0.5 and 0.8, respectively. Asterisk (*) indicates $p < 0.05$ (Bonferroni-corrected).

Focal group		DIF favoring focal		DIF favoring White Men	
Race	Gender	Large	Medium	Medium	Large
White	Women			12 , 14 , 23	
Asian	Women			6 , 12 , 14 , 23 , 27	
Black	Women	15 , 29	2 , 4 , 17	7 , 8 , 10, 12, 24	6 , 14 , 23 , 27
Hispanic	Women	15	17 , 29	10, 12, 14 , 23	27
White Hispanic	Women		29	13 , 14 , 22 , 23 , 24 , 27	6 , 10 , 12
Asian	Men				
Black	Men	29	17	10	
Hispanic	Men	29	4, 15, 17	10	
White Hispanic	Men				

B. IRT based DIF

Table VII shows the 16 items on the FCI that had a medium-to-large DIF when comparing White men to each group while accounting for educational debts as measured by the FCI post-test that we listed in Table V. All 16 were statistically significant in at least one comparison, and we report all items that had medium-to-large DIF to examine trends across groups and for transparency. For comparison with Traxler *et al.* [12], Table X lists the item difficulty parameters, accounting for educational debt. Table XI lists the ESSD values and associated standard errors and Table XII lists the infit and outfit statistics for each item.

Similar to the CTT results, most items with large or medium DIF favoring White men (11 of 16) only occurred in the comparisons to groups of women; items 12, 14, and 23 in all five comparisons and item 27 in four comparisons. Items 6 and 10 occurred in three comparisons whereas they occurred in four comparisons in the CTT results. The comparisons to Asian, Black, Hispanic, and White Hispanic men only identified item 10 favoring White men with a medium sized DIF in two cases.

Of the 11 items showing DIF favoring White men against at least one group, 10 of the 11 items (6, 7, 8, 10, 12, 14, 22, 23, 24, and 27) belonged to the first law or second law factors in the EW5 model, see Fig 2. Item 13 was the only exception and belonged to the identify forces factor. Of the five items showing DIF favoring at least one minoritized group compared to White men (items 2, 4, 15, 17, and 29), none belonged to the firstlaw or second law factors, see Fig 2. Items with DIF favoring White men almost always came from those two factors, with the exception of items 9 and 21 from the second law factor, which tended to have a small DIF favoring the minoritized groups.

The number of items exhibiting DIF was related to the size of the educational debts. This relationship, however, was less pronounced than the relationship between DIF and gender. The smallest educational debts were owed to Asian women and men, White women, and White Hispanic men.

No items with moderate or large DIF favoring White men occurred in the comparisons to Asian and White Hispanic Men while three and five items with moderate DIF occurred in the comparisons to White and Asian women. Similarly, society owed similarly large educational debts to Black women and men. The comparison to Black men, however, identified one item with moderate DIF while that to Black women identified nine items with moderate or large DIF.

VIII. DISCUSSION

Both the CTT and IRT DIF analyses yielded similar results in that both identified many items with DIF in favor of White men, and more of these items were detected when comparing White men to groups of women than to groups of minoritized men. Items 6, 12, 14, 23, and 27 were detected as having DIF favoring White men over multiple minoritized groups in both analyses, see Tables VI and VII. The size and direction of the DIF detected on all items were also similar in both analyses, see Figs. 1 and 2. The CTT indicated many more items with large DIF. This difference between the CTT and IRT comparisons likely resulted from the two methods using different scales ($\Delta\alpha_{MH}$ vs ESSD) with different cutoffs for medium and large DIF, as Figs. 1 and 2 are very similar. The results, and two figures, from the two methods also differ because the anchoring method in the IRT analysis fixed three of the ESSD values to zero in each comparison.

The results indicate an intersectional relationship between DIF and social identity groups as both IRT and CTT identified far more items with DIF in the comparisons to women than in the comparisons to minoritized men. Neither the IRT nor CTT analysis identified any items with large DIF favoring White men over other groups of men. In contrast, the CTT and IRT analyses identified many items with large DIF for comparisons of White men to groups of women. While society owed larger educational debts to Asian, Hispanic, White, and White Hispanic women than men of the same race—the educational debts for Black

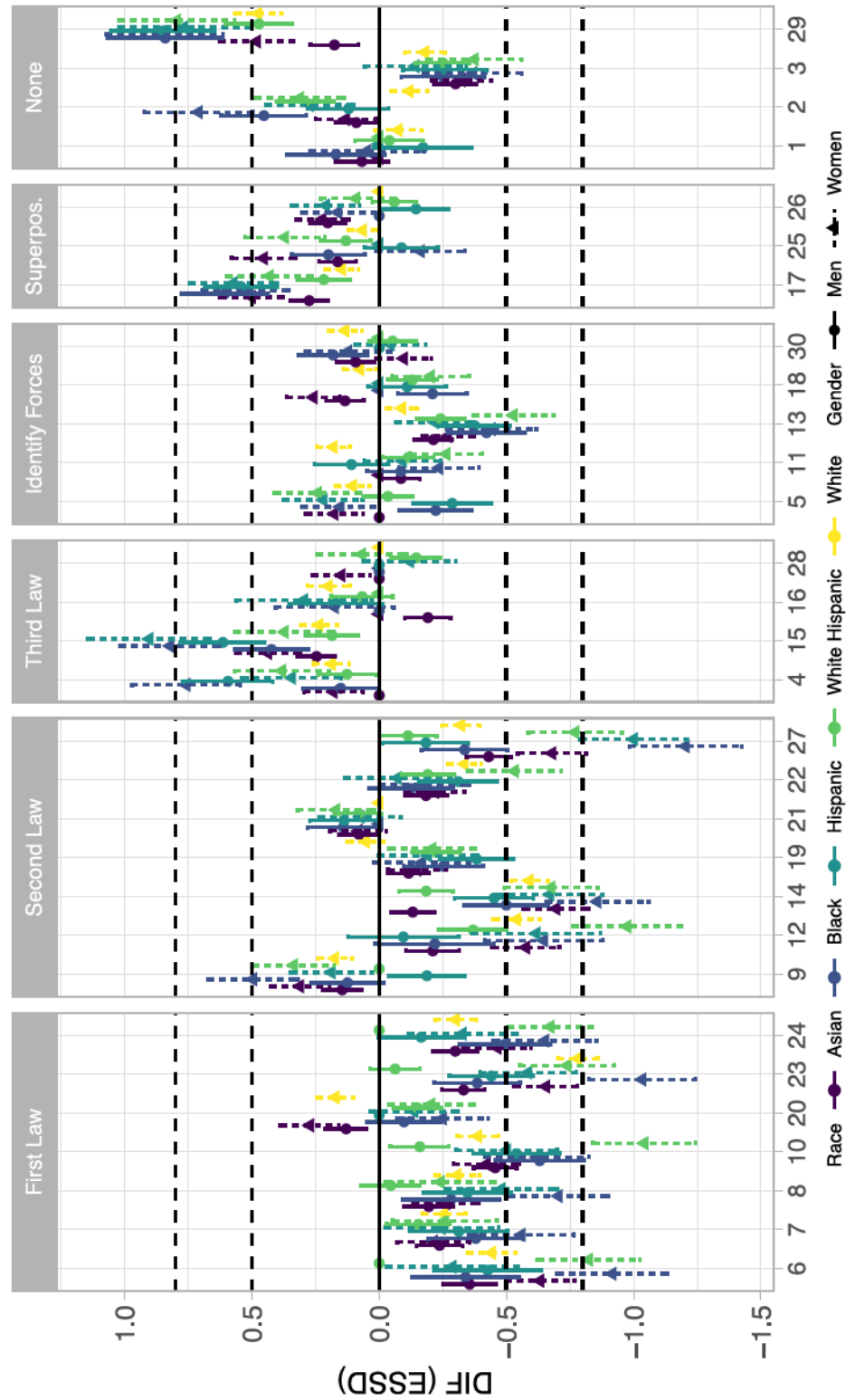


FIG. 2. DIF of the 30 FCI items using IRT accounting for educational debts organized by the factors of the EW5 model. Error bars represent 1 standard error. Items without error bars (with an ESSD of 0) were used as anchors. Negative values indicate an effect that advantages White men. Dashed lines indicate cutoffs for medium (0.5) and large (0.8) DIF values [18].

women and men were similar—the gender differences in the DIF analyses do not appear to be driven by the larger educational debts owed to women than to men. The educational debts owed to women ranged from 0.46 to 0.98 and those owed to men ranged from 0.43 to 0.98. Society owed White and Asian women similar educational debts to those owed to Asian and White Hispanic men (0.43–0.56). Yet, the DIF analysis identified many more items with DIF in the comparisons to Asian and White women. Similarly, society owed Black women and men the same educational debts; yet, the DIF analyses identified far more items with moderate or large DIF in the comparisons to Black women. Within the comparisons to women, however, intersectionality was apparent as more items favored White men in the comparisons to groups owed larger educational debts: Black, Hispanic, and White Hispanic women than White and Asian women. As shown in Traxler *et al.* [12] the DIF of items on the FCI is a gendered phenomenon. These results, however, indicate DIF is also an intersectional phenomenon with more items with larger DIF for Black, Hispanic, and White Hispanic women; the groups least represented in our data.

Traxler *et al.* [12] identified 11 items they recommended removing from the FCI: items **6, 9, 12, 14, 15, 21, 22, 23, 24, 27**, and 29. Eight of these items, in bold, favored men, 2 items favored women, and item 29 had poor psychometric properties. This list largely overlaps with the items with large DIF identified in our CTT analysis favoring White men: 6, 8, 10, 12, 14, 23, 24, and 27. Items 8 and 10 do not appear on their list and item 21 does not appear on our list. Both Traxler *et al.* [12] and our analysis identified items 9 and 15 as having DIF favoring women. Our CTT analysis, however, also identified items 2, 4, 17, 18, 25, and 26 as having large DIF favoring a group of women and item 28 as favoring Black men. We are interpreting these items as an example of artificial DIF whereby these are the items with the smallest performance differences and the large differences and DIF in favor of White men on the other items drives the artificial DIF on these items. The DIF scores alone do not indicate bias on these items [35].

Traxler *et al.* [12] found that removing 10 of those 11 items halved the difference between men and women on the post-test from $d = 0.46$ to $d = 0.23$; they did not remove item 29 in this analysis. We, however, found a much smaller shift in the educational debts owed to women when we removed the 11 items, see Table V, ranging from 16% to 3% decreases. Traxler and colleagues' recommendation to eliminate 11 items from the FCI may represent an instance of the "Golden Rule law suit settlement" [35,36]. The Golden Rule procedure attempted to minimize the score differences between Black and White test takers by choosing only those test questions that minimized the empirical performance difference between these two groups. Holland [35] applies this rule to show how to construct tests that maximize or minimize the differences

between two groups. In so doing, Holland demonstrates that DIF is neither necessary nor sufficient for determining if a test or test items are fair. DIF may identify unfair items and DIF can also identify real differences. Unfair, biased tests can have minimal DIF so long as all of the items are similarly unfair toward one group. For example, researchers can design a test that maximizes the difference in scores between two groups by creating a short assessment made of only items from a longer assessment that favored a particular group. This would be the opposite of Traxler and colleagues' [12] recommendation.

Researchers designed the Force and Motion Conceptual Evaluation (FMCE) [81], like the FCI, to assess Newtonian concepts. This similarity allows contrasting the two tests to better understand the DIF items on the FCI given that Henderson *et al.* [13] found only one item with DIF on the FMCE. One motivation for our study is that prior research has found very similar differences in scores across races and genders between the FCI and FMCE [82]. Henderson *et al.* [13] and Traxler *et al.* [12] also found similar gender differences in scores on the FMCE ($d = 0.43$ and 0.53) [83] and complete FCI post-tests ($d = 0.46$, 0.56 , and 0.69). Given the FMCE and FCI largely measure the same Newtonian concepts [84], if the FCI has a large and systematic bias against women and the FMCE does not, the FCI should report much larger gender differences in scores than the FMCE. Madsen *et al.* [85] investigated gender differences on several instruments and did not note a large difference between the FCI and FMCE. These similarities in gender differences in scores between the two instruments indicate the FCI is not biased against women.

One further difference between Traxler *et al.* [12] and our study is the development of the Eaton and Whilloughby five-factor (EW5) model for the FCI [51]. Traxler *et al.* [12] cite findings that the FCI lacks a consistent factor structure [86]. Eaton [14] contrasted several factor structures and identified their EW5 model as fitting the data well. They also confirmed that this factor structure was the same, invariant, across gender [14]. Morley *et al.* [19] extended evidence for the measurement invariance of the FCI across the ten social identity groups in this study using a dataset from LASSO that overlapped with the one used for this study. The DIF patterns on the FCI post-test align with the EW5 factor structure. All items found to exhibit large DIF favoring White men in any comparison belong to the first law and second law factors from the EW5 model. This trend is shown in Figs. 1 and 2 where all of the items below the lowest dashed lines, i.e., favoring White men, belonged to the first law or second law factors. Eaton [14] found that instruction largely maintained preexisting gender differences on these two factors while the differences decreased on the other factors. Almost all of the items with large DIF favoring the minoritized group did not belong to those two factors; the exception was item 9. These DIF patterns likely represent real

differences in physics knowledge between White men and other groups, real educational debts that society owes to those groups. Real educational debts also appear in participation rates on the AP physics exams [87]. Krakehl and Kelly [87] argue these inequities result from structural inequities, such as school funding and physics course availability and quality; stereotypes that privilege White men in physics; and a culture of marginalization in physics, for example [29,30,88].

Accounting for the educational debts society owed minoritized groups allowed us to consider for the average shift in performance on items. Our data, however, show variation in inequities across constructs on the FCI. So, while many items had medium-to-large DIF, our findings indicate this is a function of unequal distribution of educational debt across those constructs rather than item bias. Differences in group performances on the FCI are not a result of biased FCI items but represent real differences in physics knowledge between White men and minoritized groups, real educational debts that society owes to those groups. Developing, implementing, and assessing educational strategies to repay these debts owed by society requires accurate measurement of these inequities.

IX. CONCLUSIONS

DIF is neither necessary nor sufficient to claim that an item or a test is unfair or biased. Based on the results in this paper and prior research, the DIF on the FCI post-tests results from physics instruction that maintains the educational debts owed to women primarily on Newton's first and second law and not from gender bias of the FCI. The DIF results in this study and in Traxler *et al.* [12] largely overlapped. Applying their recommendation to remove 11 items, however, had little impact on the size of the educational debts measured by the FCI. If the FCI were gender biased, then we would also expect it to yield much larger differences in scores than the FMCE, but this is not the case [12,13,82,85]. The items with moderate and large DIF favoring White men almost all loaded on the two factors for Newton's first and second law. This factor structure is also invariant across gender [14] and the ten social identity groups in this study [19]. This factor invariance also implies the FCI overall makes similar measurements across social identity groups. The consistent loading on Newton's first and second law factors indicates educational debts concentrated in these content areas.

The way that we structure our physics courses, the weighting of the components of the course grade [5,89,90], the use of learning assistants [9], and numerous other aspects of the learning environment [91–93] can act to reinforce or oppose existing inequities. The DIF for items on Newton's first and second law combined with Eaton's [14] finding that instruction maintains educational debts on these factors indicates courses may need to cover this material in greater depth to repay these educational debts.

Repaying educational debts requires accurate and unbiased measurements, whether qualitative or quantitative, to guide the development and testing of interventions or instructional transformations that promote equitable instruction. These assessments can adopt modern psychometric techniques, such as cognitive diagnostic models [94], that provide more fine-grained, actionable measures and flexibility in timing and content coverage than fixed length tests to better meet instructors needs. Constructing new assessments within a framework, such as evidence centered design [95–97], can ensure that they meet instructors and researchers needs while providing accurate and unbiased measurements.

ACKNOWLEDGMENTS

This work was funded in part by NSF-IUSE Grants No. DUE-1928596. This work was completed while coauthor Robert M. Talbot was serving at the National Science Foundation.

APPENDIX: IRT-BASED DIF WITHOUT ACCOUNTING FOR EDUCATIONAL DEBT

For transparency and comparison with the IRT results presented in Sec. VII B, we present results from an IRT analysis using the Rasch model where we do not account for educational debts. This necessarily changes the anchoring scheme used in the analysis. Anchor selection proceeds identically to that described in Sec. VI B 2 up through the analysis where the preliminary item difficulties are extracted. At this point, we compare the item difficulties across the groups and select the three items with the closest difficulties across groups as anchors. This acts to shift the ESSD of each item to have larger values favoring White men, which can be seen in the different range of ESSD values in Fig. 2, approximately 1 to -1 , and Fig. 3, approximately 0 to -2 . The detailed output of the Rasch model when educational debt is not accounted for is presented in Tables XIII, XIV, and XV which list the item difficulty parameters, ESSD and associated standard errors, and infit and outfit statistics for each item, respectively.

Table VIII shows that 25 of the 30 items on the FCI had a medium-to-large ESSD when comparing White men to any other group, while not accounting for educational debts. Of these, 24 were statistically significant in at least one comparison, but we report all items that had medium-to-large effect sizes to examine trends across groups and for transparency.

Similar to the CTT and previous IRT analysis, these results show that White men are favored by questions on the FCI. No questions had a medium or large ESSD in favor of the focal group (i.e., against White men). Of the 25 items found to favor White men (1, 3, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 16, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, and 30), one (item 21) was only flagged when comparing White

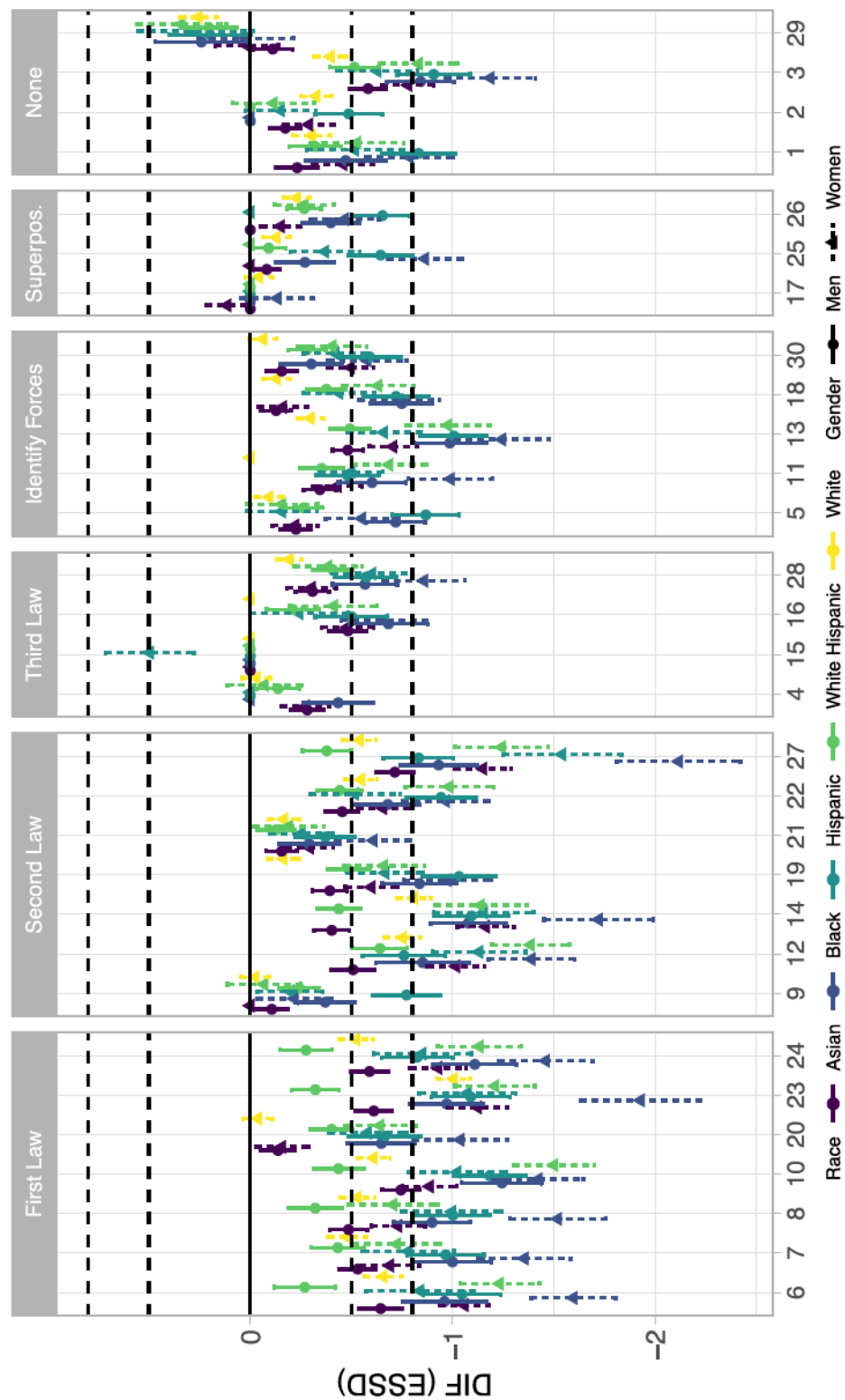


FIG. 3. ESSD effect sizes, not accounting for educational debts, for the DIF of the 30 FCI items organized by the factors of the EW5 model [18]. Error bars represent 1 standard error. Items without error bars (with an ESSD of 0) were used as anchors. Negative values indicate an effect that advantages White men. Dashed lines indicate cutoffs for medium (0.5) and large (0.8) ESSD values.

TABLE VIII. Item DIF across race or ethnicity and gender using Item Response Theory. No educational debts and three anchors. Cutoffs for medium and large were ESSD scores of 0.5 and 0.8, respectively. Asterisk (*) indicates $p < 0.05$ (Bonferroni-corrected).

Focal group		DIF favoring focal		DIF favoring White Men	
Race	Gender	Large	Medium	Medium	Large
White	Women	...	...	6, 8, 10, 12, 22, 24, 27	14, 23
Asian	Women	...	...	3, 7, 8, 13, 19, 22	6, 10, 12, 14, 23, 24, 27
Black	Women	...	...	1, 5, 16, 18, 21, 30	3, 6, 7, 8, 10, 11, 12, 13, 14, 19, 20, 22, 23, 24, 25, 27, 28
Hispanic	Women	...	...	1, 3, 7, 13, 19, 20, 22, 28	6, 8, 10, 12, 14, 23, 24, 27
White Hispanic	Women	...	...	1, 7, 8, 11, 18, 19, 20	3, 6, 10, 12, 13, 14, 22, 23, 24, 27
Asian	Men	...	...	3, 6, 7, 10, 12, 23, 24, 27	...
Black	Men	...	...	5, 11, 16, 18, 20, 22, 28	3, 6, 7, 8, 10, 12, 13, 14, 19, 23, 24, 27
Hispanic	Men	...	...	9, 12, 16, 18, 20, 25, 26, 28, 30	1, 3, 56, 7, 8, 10, 13, 14, 19, 22, 23, 24, 27
White Hispanic	Men	...	...	3, 12	...

men to Black women, and two (items 9 and 26) were only flagged when comparing White men with Hispanic men. Item 12 showed up in all 9 comparisons, and items 3, 6, 10, 23, 24, and 27 showed up in 8 comparisons (all but White Hispanic men). Items 7, 8, 14, and 22 show up in 7 comparisons with 8, 14, and 22 not showing up for Asian men and White Hispanic men, and 7 not showing up for White women and White Hispanic men. This is consistent with our CTT and prior IRT findings, where items 6, 10, 12, 14, 23, and 27 consistently favored White men over women. This analysis extends this earlier finding to favor White men over minoritized groups of men. Items 3, 7, and 8 were not frequently flagged in the prior CTT and IRT analyses, which is consistent with these items often being flagged as having a Medium ESSD (4 of 9 comparisons for item 3, 4 of 9 comparisons for item 7, and 3 of 9 comparisons for item 8, Table VIII).

The total number of items identified as having medium-to-large ESSD in favor of White men did not differ greatly across genders. Within most races, however, more items tended to have a large ESSD favoring White men compared to women than men of that race. Figure 3 shows this trend in the much lower ESSD scores for women, shown with triangles, than men, shown with circles, of the same race. Drawing on Table VIII to look at the distribution of medium and large ESSD values for men and women of each race. Seven of 13 items had a large ESSD against Asian women (54%), compared to 0 of 8 for Asian men (0%). This comparison is 17 of 23 items for Black women (74%), compared to 12 of 19 for Black men (63%), and 10

of 17 items for White Hispanic women (59%) compared to 0 of 2 for White Hispanic men (0%). The exception to this trend was Hispanic women, where 8 of 16 items had a large ESSD against Hispanic women (50%) compared to 14 of 23 for Hispanic men (61%). This one exception contradicts our findings in the CTT and IRT analysis that accounts for educational debt, where more items favored White men over Hispanic women than over Hispanic men.

Figure 3 shows the ESSD effect sizes grouped by the EW5 factors. Comparing with Fig. 2, we see that the overall pattern in effect sizes is similar to when educational debts are accounted for, except that in this case all of the effect sizes are shifted down such that all items either increase the advantage White men have on that item or decrease their disadvantage. The first law and second law factors consistently have ESSD effect sizes that favor White men more than the third law, identify forces, and superposition factors. The exceptions to this trend are also consistent, with items 9 and 21 in the second law having a higher ESSD value than the rest of the times in the second law factor, similar to Fig. 2.

Figure 3 also shows that a small number of items were frequently chosen as anchors, item 2 (3 times), item 4 (3 times), item 15 (8 times), and item 17 (6 times). These items all showed up as favoring the focal (minoritized) group in the IRT analysis which accounted for educational debts. The only exception to this is item 29, which showed up as favoring the focal group but was explicitly excluded from being an anchor due to prior studies showing its bad fit.

TABLE IX. $\Delta\alpha_{MH}$ and standard errors.

Factor	Item	Asian				Black				Hispanic				White Hispanic				White	
		Men		Women		Men		Women		Men		Women		Men		Women		Women	
		$\Delta\alpha_{MH}$	SE	$\Delta\alpha_{MH}$	SE	$\Delta\alpha_{MH}$	SE	$\Delta\alpha_{MH}$	SE	$\Delta\alpha_{MH}$	SE	$\Delta\alpha_{MH}$	SE	$\Delta\alpha_{MH}$	SE	$\Delta\alpha_{MH}$	SE	$\Delta\alpha_{MH}$	SE
First law	6	-0.60	0.42	-1.39	0.70	-0.11	0.58	-1.96	0.68	-0.45	0.49	-0.28	0.63	0.26	0.44	-1.94	0.68	-1.13	0.41
First law	7	-0.32	0.45	-0.32	0.49	-0.54	0.52	-1.26	0.60	-0.30	0.41	-0.55	0.70	-0.34	0.34	-0.70	0.50	-0.65	0.43
First law	8	-0.31	0.41	-0.46	0.57	-0.32	0.46	-1.62	0.57	-0.37	0.42	-1.28	0.73	-0.02	0.30	-0.60	0.62	-0.83	0.40
First law	10	-0.86	0.44	-0.73	0.70	-0.62	0.57	-1.22	0.66	-0.49	0.49	-1.05	0.69	-0.07	0.33	-2.66	0.68	-1.06	0.53
First law	20	0.88	0.42	1.40	0.53	0.25	0.46	-0.54	0.53	0.39	0.40	-0.36	0.61	-0.36	0.32	-0.51	0.57	0.58	0.42
First law	23	-0.76	0.42	-1.91	0.48	-0.47	0.50	-2.58	0.67	-0.43	0.35	-1.95	0.74	0.08	0.32	-1.95	0.77	-2.29	0.45
First law	24	-0.34	0.45	-0.82	0.67	-0.15	0.61	-1.28	0.66	0.51	0.45	-0.29	0.67	0.49	0.38	-1.66	0.70	-0.77	0.51
Second law	9	0.99	0.44	1.36	0.52	0.86	0.57	1.58	0.63	-0.14	0.34	1.07	0.67	0.13	0.28	1.03	0.77	0.64	0.45
Second law	12	-0.21	0.42	-1.32	0.56	-0.06	0.47	-1.39	0.63	0.21	0.54	-1.47	0.67	-1.03	0.37	-2.40	0.62	-1.42	0.38
Second law	14	-0.26	0.38	-2.13	0.57	-1.27	0.56	-2.33	0.59	-0.88	0.43	-2.40	0.70	-0.61	0.26	-2.01	0.67	-1.78	0.38
Second law	19	0.07	0.41	-0.06	0.48	0.12	0.47	-0.14	0.57	-0.35	0.37	-0.33	0.66	-0.49	0.32	-0.38	0.56	0.33	0.50
Second law	21	0.44	0.34	0.46	0.49	0.27	0.48	-0.04	0.52	0.40	0.38	0.06	0.70	0.06	0.30	0.15	0.63	-0.13	0.31
Second law	22	-0.41	0.38	-0.58	0.51	-0.13	0.40	-0.38	0.49	-0.36	0.42	-0.33	0.56	-0.49	0.30	-1.44	0.68	-0.92	0.42
Second law	27	-1.15	0.36	-1.95	0.57	-0.55	0.55	-2.86	0.58	-0.11	0.38	-3.86	0.87	-0.28	0.27	-2.08	0.56	-0.90	0.40
Third law	4	0.45	0.37	1.06	0.56	0.97	0.47	1.92	0.49	1.63	0.38	1.93	0.60	0.62	0.33	1.46	0.55	0.77	0.41
Third law	15	0.73	0.27	1.10	0.40	0.67	0.44	1.54	0.47	0.98	0.34	2.28	0.60	0.19	0.26	0.55	0.47	0.54	0.32
Third law	16	-0.12	0.41	0.45	0.58	0.58	0.47	0.52	0.64	0.84	0.41	1.38	0.77	0.41	0.33	0.30	0.65	0.69	0.46
Third law	28	0.82	0.49	1.38	0.65	1.62	0.67	1.14	0.57	1.10	0.48	0.24	0.69	0.11	0.36	0.83	0.79	0.30	0.49
Identify forces	5	0.34	0.34	0.85	0.54	-0.64	0.48	0.83	0.51	-0.38	0.33	1.07	0.61	0.05	0.31	0.92	0.59	0.44	0.41
Identify forces	11	0.09	0.33	0.49	0.50	0.41	0.54	-0.38	0.58	0.98	0.42	0.06	0.56	-0.15	0.33	-0.74	0.62	0.69	0.45
Identify forces	13	-0.18	0.42	-0.26	0.74	-0.19	0.52	-0.69	0.71	0.29	0.48	-0.01	0.83	-0.14	0.37	-1.16	0.87	-0.05	0.54
Identify forces	18	1.02	0.41	1.65	0.59	0.11	0.49	0.72	0.57	0.57	0.38	0.57	0.74	0.03	0.33	-0.25	0.73	0.51	0.46
Identify forces	30	0.66	0.38	0.01	0.55	1.23	0.51	0.81	0.59	0.54	0.36	0.06	0.61	0.18	0.33	0.23	0.71	0.62	0.42
Superposition	17	1.07	0.32	1.79	0.45	1.73	0.43	1.69	0.51	1.67	0.41	2.09	0.66	0.62	0.31	1.22	0.56	0.53	0.40
Superposition	25	0.81	0.33	1.77	0.48	0.72	0.44	-0.25	0.54	0.22	0.35	-0.20	0.62	0.61	0.29	1.38	0.52	0.47	0.46
Superposition	26	1.49	0.45	1.46	0.73	0.91	0.62	1.64	0.64	0.66	0.55	2.00	0.83	0.01	0.36	0.86	0.78	0.12	0.55
None	1	0.58	0.36	0.42	0.47	0.73	0.44	0.09	0.47	-0.09	0.36	0.38	0.63	-0.14	0.32	-0.15	0.59	-0.13	0.42
None	2	0.41	0.30	0.49	0.47	1.05	0.45	1.64	0.44	0.37	0.35	0.79	0.59	0.72	0.26	0.65	0.60	-0.36	0.39
None	3	-0.74	0.35	-0.81	0.58	-0.49	0.42	-0.93	0.57	-0.40	0.41	-0.55	0.67	-0.82	0.33	-1.08	0.62	-0.50	0.44
None	29	0.63	0.38	1.28	0.46	1.65	0.52	1.36	0.59	1.56	0.44	1.74	0.66	0.88	0.35	1.57	0.62	1.01	0.35

TABLE X. Item difficulties from the Rasch models. Educational debts are accounted for.

Factor	Item	Asian		Black		Hispanic		White Hispanic		White	
		Men	Women	Men	Women	Men	Women	Men	Women	Men	Women
First law	6	-2.51	-2.18	-2.65	-2.07	-2.58	-2.66	-2.99 [†]	-2.09	-2.99	-2.46
First law	7	-1.98	-2.03	-1.90	-1.74	-2.00	-1.99	-2.11	-2.02	-2.30	-2.00
First law	8	-1.73	-1.65	-1.70	-1.27	-1.65	-1.34	-1.94	-1.73	-2.00	-1.62
First law	10	-1.53	-1.61	-1.49	-1.54	-1.63	-1.50	-1.97	-1.00	-2.17	-1.70
First law	20	-1.16	-1.35	-0.84	-0.68	-0.96 [†]	-0.74	-0.75	-0.71	-0.96	-1.18
First law	23	-0.78	-0.31	-0.82	-0.11	-0.81	-0.33	-1.20	-0.39	-1.28	-0.26
First law	24	-1.68	-1.47	-1.56	-1.43	-1.93	-1.66	-2.10 [†]	-1.34	-2.10	-1.73
Second law	9	-0.58	-0.82	-0.52	-0.93	-0.12	-0.68	-0.35 [†]	-0.77	-0.35	-0.58
Second law	12	-2.53	-2.07	-2.59	-2.17	-2.72	-2.05	-2.36	-1.73	-2.81	-2.16
Second law	14	-0.84	0.00	-0.41	-0.05	-0.55	0.12	-0.79	-0.21	-1.05	-0.26
Second law	19	-1.19	-1.14	-1.06	-1.18	-0.95	-1.07	-1.06	-1.11	-1.36	-1.42
Second law	21	0.41	0.41	0.33	0.55 [†]	0.37	0.40	0.42	0.31	0.54	0.54 [†]
Second law	22	-0.74	-0.69	-0.86	-0.82	-0.68	-0.90	-0.75	-0.36	-1.02	-0.57
Second law	27	-0.86	-0.52	-1.11	-0.17	-1.31	0.12	-1.36	-0.60	-1.50	-1.10
Third law	4	-1.46 [†]	-1.71	-1.63	-2.24	-2.06	-1.93	-1.63	-1.89	-1.46	-1.70
Third law	15	-0.78	-1.05	-0.94	-1.32	-1.08	-1.77	-0.66	-0.86	-0.39	-0.71
Third law	16	-1.83	-2.10 [†]	-2.09 [†]	-2.27	-2.26	-2.46	-2.18	-2.10 [†]	-2.10	-2.33
Third law	28	-1.44 [†]	-1.65 [†]	-1.45	-1.44 [†]	-1.45 [†]	-1.27	-1.25	-1.52	-1.45	-1.45 [†]
Identify forces	5	-0.23 [†]	-0.50	0.09	-0.42	0.12	-0.62	-0.18	-0.54	-0.23	-0.37
Identify forces	11	-0.41	-0.55 [†]	-0.44	-0.27	-0.68	-0.39	-0.38	-0.21	-0.55	-0.79
Identify forces	13	-0.60	-0.52	-0.39	-0.42	-0.51	-0.56	-0.59	-0.27	-0.93	-0.81
Identify forces	18	-0.97	-1.14	-0.49	-0.76 [†]	-0.64	-0.76 [†]	-0.57	-0.51	-0.76	-0.86
Identify forces	30	-0.34	-0.04	-0.44	-0.34	-0.19 [†]	-0.11	-0.11	-0.19 [†]	-0.19	-0.38
Superposition	17	-0.21	-0.54	-0.59	-0.42	-0.41	-0.79	-0.10	-0.33	0.25	0.04
Superposition	25	-0.29	-0.73	-0.31	0.18	0.08	-0.03 [†]	-0.23	-0.51	-0.03	-0.12
Superposition	26	0.27	0.25	0.62 [†]	0.40	0.81	0.18	0.72	0.49	0.63	0.63 [†]
None	1	-2.45	-2.36 [†]	-2.53	-2.41	-2.19	-2.36 [†]	-2.31	-2.36 [†]	-2.36	-2.27
None	2	-0.94	-0.99	-1.34	-1.57	-0.93	-1.23	-1.20	-1.17	-0.80	-0.64
None	3	-1.05	-1.03	-1.20	-1.09	-1.22	-1.28	-1.16	-1.06	-1.49	-1.26
None	29	-1.99	-2.37	-2.62	-2.58	-2.58	-2.72	-2.33	-2.61	-1.74	-2.32

TABLE XI. ESSD values and standard errors, accounting for educational debts.

Factor	Item	Asian				Black				Hispanic				White Hispanic				White			
		Men		Women		Men		Women		Men		Women		Men		Women		Men		Women	
		ESSD	SE	ESSD	SE	ESSD	SE	ESSD	SE	ESSD	SE	ESSD	SE	ESSD	SE	ESSD	SE	ESSD	SE	ESSD	SE
First law	6	-0.45	0.11	-0.47	0.13	-0.48	0.21	-0.66	0.21	-0.37	0.22	0.00	0.00	0.14	0.15	-0.56	0.19	-0.19	0.09	-0.19	0.09
First law	7	-0.33	0.10	0.00	0.00	-0.52	0.19	-0.29	0.20	-0.26	0.20	0.05	0.19	-0.04	0.14	0.00	0.00	0.00	0.00	0.00	0.00
First law	8	-0.29	0.10	-0.10	0.13	-0.42	0.19	-0.44	0.20	-0.30	0.19	-0.20	0.20	0.07	0.12	0.00	0.00	-0.07	0.08	-0.07	0.08
First law	10	-0.55	0.09	-0.26	0.13	-0.77	0.18	-0.36	0.20	-0.49	0.18	-0.21	0.18	-0.05	0.13	-0.77	0.18	-0.14	0.08	-0.14	0.08
First law	20	0.05	0.08	0.41	0.11	-0.21	0.15	0.00	0.00	0.00	0.00	0.09	0.17	-0.05	0.10	0.04	0.17	0.39	0.07	0.39	0.07
First law	23	-0.42	0.09	-0.50	0.11	-0.51	0.17	-0.78	0.21	-0.39	0.16	-0.34	0.17	0.04	0.11	-0.48	0.17	-0.54	0.07	-0.54	0.07
First law	24	-0.39	0.10	-0.30	0.13	-0.63	0.18	-0.39	0.19	-0.11	0.18	0.00	0.00	0.12	0.13	-0.40	0.17	0.00	0.00	0.00	0.00
Second law	9	0.07	0.08	0.45	0.12	0.00	0.00	0.72	0.19	-0.15	0.14	0.39	0.16	0.08	0.09	0.56	0.18	0.39	0.07	0.39	0.07
Second law	12	-0.31	0.11	-0.41	0.14	-0.36	0.24	-0.38	0.23	0.00	0.00	-0.30	0.23	-0.25	0.14	-0.71	0.19	-0.30	0.09	-0.30	0.09
Second law	14	-0.22	0.08	-0.55	0.12	-0.62	0.17	-0.61	0.20	-0.40	0.16	-0.43	0.19	-0.08	0.10	-0.43	0.17	-0.36	0.08	-0.36	0.08
Second law	19	-0.20	0.09	0.00	0.00	-0.38	0.16	0.08	0.18	-0.33	0.15	0.06	0.17	-0.12	0.10	0.04	0.17	0.28	0.07	0.28	0.07
Second law	21	0.00	0.00	0.21	0.10	0.05	0.14	0.22	0.17	0.17	0.15	0.25	0.18	0.16	0.09	0.37	0.17	0.22	0.07	0.22	0.07
Second law	22	-0.27	0.09	-0.08	0.11	-0.24	0.17	0.05	0.18	-0.26	0.15	0.16	0.19	-0.09	0.10	-0.29	0.18	-0.11	0.07	-0.11	0.07
Second law	27	-0.52	0.10	-0.52	0.12	-0.46	0.18	-0.94	0.22	-0.13	0.17	-0.72	0.19	0.00	0.00	-0.51	0.18	-0.09	0.08	-0.09	0.08
Third law	4	-0.09	0.09	0.33	0.12	0.00	0.00	1.01	0.23	0.64	0.19	0.60	0.21	0.23	0.12	0.64	0.19	0.42	0.07	0.42	0.07
Third law	15	0.17	0.08	0.57	0.13	0.32	0.15	1.05	0.21	0.66	0.18	1.13	0.28	0.28	0.10	0.60	0.20	0.45	0.07	0.45	0.07
Third law	16	-0.28	0.09	0.16	0.14	-0.19	0.18	0.44	0.22	0.22	0.20	0.58	0.27	0.18	0.13	0.33	0.19	0.44	0.09	0.44	0.09
Third law	28	-0.12	0.08	0.30	0.12	-0.10	0.13	0.20	0.18	0.12	0.16	0.13	0.18	-0.04	0.10	0.32	0.18	0.25	0.07	0.25	0.07
Identify forces	5	-0.05	0.08	0.31	0.11	-0.32	0.14	0.36	0.17	-0.25	0.15	0.42	0.17	0.06	0.09	0.46	0.16	0.31	0.07	0.31	0.07
Identify forces	11	-0.17	0.08	0.12	0.10	-0.19	0.13	0.00	0.00	0.15	0.15	0.12	0.16	0.00	0.00	0.00	0.00	0.40	0.07	0.40	0.07
Identify forces	13	-0.30	0.08	-0.13	0.09	-0.54	0.16	-0.21	0.15	-0.33	0.14	0.00	0.00	-0.14	0.09	-0.28	0.15	0.13	0.06	0.13	0.06
Identify forces	18	0.05	0.08	0.40	0.11	-0.32	0.15	0.25	0.18	-0.07	0.14	0.20	0.13	-0.03	0.09	0.03	0.16	0.29	0.07	0.29	0.07
Identify forces	30	0.01	0.08	0.00	0.00	0.08	0.13	0.33	0.17	0.00	0.00	0.15	0.17	0.04	0.08	0.21	0.15	0.34	0.07	0.34	0.07
Superposition	17	0.20	0.08	0.63	0.12	0.51	0.16	0.73	0.17	0.58	0.15	0.77	0.21	0.30	0.10	0.64	0.17	0.35	0.07	0.35	0.07
Superposition	25	0.09	0.08	0.59	0.12	0.10	0.15	0.00	0.00	-0.05	0.14	0.18	0.17	0.22	0.09	0.59	0.17	0.27	0.07	0.27	0.07
Superposition	26	0.13	0.07	0.35	0.10	0.00	0.00	0.35	0.15	-0.11	0.13	0.39	0.15	0.00	0.00	0.29	0.15	0.15	0.06	0.15	0.06
None	1	0.00	0.00	0.19	0.15	0.03	0.20	0.31	0.22	-0.12	0.20	0.33	0.23	0.08	0.13	0.22	0.21	0.17	0.09	0.17	0.09
None	2	0.00	0.00	0.27	0.12	0.34	0.17	0.95	0.21	0.16	0.17	0.50	0.18	0.39	0.11	0.55	0.19	0.10	0.07	0.10	0.07
None	3	-0.39	0.09	-0.17	0.12	-0.37	0.16	-0.12	0.19	-0.21	0.16	0.11	0.19	-0.14	0.11	-0.12	0.18	0.00	0.00	0.00	0.00
None	29	0.09	0.10	0.63	0.15	0.72	0.21	1.09	0.24	0.90	0.21	1.03	0.25	0.58	0.13	1.05	0.23	0.71	0.09	0.71	0.09

TABLE XII. Infit and outfit statistics for the Rasch models, accounting for educational debt.

Factor	Item	Asian						Black						Hispanic						White Hispanic						White					
		Men			Women			Men			Women			Men			Women			Men			Women			Men			Women		
		Infit	Outfit	Infit	Outfit	Infit	Outfit	Infit	Outfit	Infit	Outfit	Infit	Outfit	Infit	Outfit	Infit	Outfit	Infit	Outfit	Infit	Outfit	Infit	Outfit	Infit	Outfit	Infit	Outfit	Infit	Outfit	Infit	Outfit
First law	6	0.99	0.91	0.86	0.78	0.93	0.99	1.00	1.02	0.94	0.78	0.97	0.74	0.87	0.63	0.92	0.76	0.97	0.72	0.93	0.81										
First law	7	0.89	0.71	1.02	0.98	1.02	1.04	0.90	0.84	0.99	0.96	0.82	0.64	0.96	0.87	1.06	1.00	0.99	0.98	1.02	1.08										
First law	8	0.97	0.87	0.97	0.95	0.99	0.94	0.95	0.89	0.98	0.95	0.95	0.82	0.95	0.84	1.17	1.18	1.00	0.89	1.05	1.01										
First law	10	0.82	0.65	0.85	0.69	0.94	0.98	0.90	0.81	0.88	0.80	0.89	0.90	0.93	0.78	0.87	0.80	0.86	0.64	0.90	0.81										
First law	20	0.96	0.93	0.95	0.87	0.92	0.88	1.05	1.05	0.98	0.99	0.95	0.87	0.91	0.85	0.98	1.04	0.95	0.92	0.97	1.01										
First law	23	0.96	0.90	0.94	0.90	1.01	0.98	0.94	0.94	1.06	1.03	0.89	0.78	0.96	0.84	0.96	0.90	0.92	0.81	0.92	0.89										
First law	24	0.91	0.86	0.96	0.86	0.96	0.98	0.98	1.04	0.94	1.02	0.99	1.06	0.94	0.97	0.91	0.83	0.84	0.61	0.96	0.84										
Second law	9	0.87	0.81	1.09	1.11	0.92	0.86	0.86	0.80	1.03	1.06	0.85	0.78	0.92	0.86	0.97	0.91	0.96	0.91	0.98	0.94										
Second law	12	0.98	0.96	0.93	0.80	1.11	1.01	0.99	1.03	0.98	1.08	1.04	1.14	0.96	0.92	0.98	1.05	1.00	1.02	1.02	1.01										
Second law	14	1.01	0.97	1.04	1.01	1.01	1.03	0.94	0.93	0.95	0.92	1.19	1.31	1.05	1.09	0.92	0.89	1.06	1.04	1.02	0.99										
Second law	19	0.92	0.89	1.00	1.03	0.83	0.77	0.97	1.01	0.89	0.84	1.05	1.05	0.87	0.79	1.00	0.97	0.89	0.76	0.91	0.87										
Second law	21	0.99	1.02	0.93	0.97	0.96	1.00	0.83	0.79	0.85	0.86	1.04	1.07	0.96	0.95	1.02	1.15	1.07	1.11	1.03	1.08										
Second law	22	1.02	0.96	1.05	1.04	1.04	1.04	0.94	0.88	0.97	0.92	1.23	1.13	0.95	0.85	0.98	1.04	0.98	0.85	1.02	1.00										
Second law	27	1.01	0.95	0.97	0.98	1.07	1.06	1.08	1.18	0.97	0.91	0.82	0.78	1.01	1.00	1.10	1.13	1.00	0.94	1.07	1.03										
Third law	4	0.84	0.89	0.86	0.77	0.89	0.82	1.05	1.13	0.99	1.04	0.88	1.22	0.93	0.84	0.84	0.84	0.95	0.88	0.91	0.80										
Third law	15	1.10	1.09	1.22	1.20	0.98	0.95	0.99	0.98	1.06	1.08	1.16	1.07	1.21	1.26	1.10	1.17	1.22	1.25	1.10	1.12										
Third law	16	0.93	0.94	0.93	0.86	0.94	0.98	1.05	1.09	0.91	0.82	0.97	1.29	0.98	1.13	0.87	0.71	0.95	0.92	0.93	0.85										
Third law	28	0.82	0.74	0.80	0.68	0.76	0.68	0.97	0.90	0.83	0.79	0.80	0.66	0.82	0.69	0.82	0.82	0.80	0.63	0.84	0.73										
Identify forces	5	0.99	1.03	0.92	0.91	0.84	0.84	0.98	0.93	0.94	1.01	0.88	0.81	0.92	0.90	0.95	0.93	0.92	0.88	0.89	0.85										
Identify forces	11	0.93	0.90	0.93	0.91	0.89	0.89	0.93	0.93	0.87	0.86	0.93	0.83	1.01	1.03	0.91	0.83	0.92	0.90	0.84	0.81										
Identify forces	13	0.82	0.77	0.73	0.67	0.86	0.83	0.85	0.81	0.77	0.74	0.81	0.78	0.77	0.69	0.78	0.70	0.74	0.63	0.74	0.66										
Identify forces	18	0.94	0.88	0.85	0.77	0.80	0.71	0.96	0.95	0.93	0.88	0.70	0.63	0.81	0.73	0.84	0.78	0.84	0.74	0.78	0.70										
Identify forces	30	1.04	1.03	0.87	0.87	0.86	0.87	1.02	1.05	0.95	0.96	0.86	0.77	0.82	0.79	0.85	0.86	0.89	0.87	0.91	0.89										
Superposition	17	0.98	1.00	1.03	0.99	1.10	1.17	0.93	0.93	0.90	0.88	1.03	1.00	1.10	1.10	0.96	0.95	0.98	1.01	0.96	0.96										
Superposition	25	0.97	0.94	1.08	1.11	1.05	1.14	0.87	0.86	0.95	0.98	1.19	1.46	0.93	0.92	0.86	0.82	0.88	0.87	0.90	0.90										
Superposition	26	0.77	0.75	0.75	0.71	0.78	0.77	0.71	0.65	0.72	0.61	0.68	0.82	0.83	0.82	0.74	0.63	0.79	0.75	0.74	0.70										
None	1	1.02	1.03	0.90	0.74	0.88	0.67	0.88	0.73	1.03	1.02	0.94	0.80	0.99	0.83	1.00	0.95	1.02	0.93	0.98	1.00										
None	2	1.06	1.02	1.02	1.02	1.06	1.02	1.01	0.93	1.09	1.07	0.89	0.76	1.11	1.07	1.08	1.03	1.07	1.00	1.04	1.02										
None	3	1.02	1.02	0.96	0.86	0.90	0.91	0.94	0.94	0.99	0.99	0.90	0.85	0.98	0.97	0.98	0.95	1.06	1.12	1.02	0.99										
None	29	1.01	1.22	1.00	1.19	1.08	1.55	0.96	1.07	1.00	1.51	1.04	1.92	1.05	1.41	1.00	1.27	1.19	1.68	1.07	1.50										

TABLE XIII. Item difficulties from the Rasch models. Educational debts are not accounted for.

Factor	Item	Asian		Black		Hispanic		White Hispanic		White	
		Men	Women	Men	Women	Men	Women	Men	Women	Men	Women
First law	6	-2.11	-1.57	-2.01	-1.25	-1.93	-2.00	-2.67	-1.58	-2.99	-2.19
First law	7	-1.58	-1.42	-1.27	-0.92	-1.35	-1.34	-1.78	-1.52	-2.30	-1.73
First law	8	-1.33	-1.04	-1.06	-0.45	-1.00	-0.68	-1.61	-1.23	-2.00	-1.35
First law	10	-1.13	-1.00	-0.85	-0.72	-0.99	-0.84	-1.64	-0.49	-2.17	-1.43
First law	20	-0.76	-0.74	-0.20	0.15	-0.26	-0.08	-0.41	-0.21	-0.96	-0.91
First law	23	-0.38	0.30	-0.18	0.71	-0.16	0.33	-0.86	0.11	-1.28	0.01
First law	24	-1.28	-0.86	-0.92	-0.60	-1.29	-1.01	-1.76	-0.84	-2.10	-1.46
Second law	9	-0.18	-0.34 [†]	0.12	-0.11	0.53	-0.02	0.01	-0.26	-0.35	-0.31
Second law	12	-2.13	-1.46	-1.95	-1.35	-2.07	-1.39	-2.03	-1.23	-2.81	-1.89
Second law	14	-0.45	0.61	0.23	0.77	0.10	0.77	-0.45	0.30	-1.05	0.01
Second law	19	-0.79	-0.53	-0.42	-0.35	-0.30	-0.41	-0.72	-0.61	-1.36	-1.15
Second law	21	0.81	1.03	0.97	1.33	1.02	1.05	0.75	0.81	0.55	0.79
Second law	22	-0.34	-0.08	-0.23	0.01	-0.03	-0.25	-0.41	0.14	-1.02	-0.30
Second law	27	-0.46	0.09	-0.47	0.65	-0.67	0.77	-1.02	-0.09	-1.50	-0.82
Third law	4	-1.06	-1.10	-0.99	-1.46 [†]	-1.46 [†]	-1.46 [†]	-1.29	-1.38	-1.46	-1.42
Third law	15	-0.39 [†]	-0.40 [†]	-0.39 [†]	-0.40 [†]	-0.40 [†]	-1.12	-0.39 [†]	-0.39 [†]	-0.40	-0.40 [†]
Third law	16	-1.43	-1.48	-1.39	-1.45	-1.61	-1.81	-1.84	-1.65	-2.09	-2.09 [†]
Third law	28	-1.00	-1.04	-0.83	-0.57	-0.87	-0.61	-0.91	-1.01	-1.44	-1.20
Identify forces	5	0.13	0.11	0.73	0.41	0.77	0.03	0.16	-0.03	-0.23	-0.10
Identify forces	11	-0.02	0.09	0.20	0.56	-0.03	0.27	-0.05	0.30	-0.54	-0.54 [†]
Identify forces	13	-0.20	0.09	0.25	0.41	0.14	0.09	-0.25	0.23	-0.93	-0.54
Identify forces	18	-0.57	-0.53	0.15	0.03	0.01	-0.08	-0.24	-0.00	-0.76	-0.59
Identify forces	30	0.06	0.57	0.20	0.48	0.49	0.55	0.23	0.33	-0.19	-0.10
Superposition	17	0.24 [†]	0.07	0.24 [†]	0.41	0.24 [†]	0.24 [†]	0.25 [†]	0.24 [†]	0.25	0.31
Superposition	25	0.10	-0.03 [†]	0.33	1.01	0.73	0.64	0.11	-0.03 [†]	-0.03	0.15
Superposition	26	0.62 [†]	0.86	1.21	1.22	1.46	0.62 [†]	1.05	0.99	0.62	0.95
None	1	-2.05	-1.79	-1.89	-1.58	-1.54	-1.75	-1.97	-1.79	-2.36	-2.00
None	2	-0.54	-0.38	-0.79 [†]	-0.80 [†]	-0.28	-0.57	-0.81 [†]	-0.66	-0.80	-0.37
None	3	-0.65	-0.42	-0.56	-0.27	-0.57	-0.63	-0.82	-0.55	-1.49	-0.99
None	29	-1.59	-1.76	-1.98	-1.76	-1.93	-2.06	-1.99	-2.11	-1.74	-2.05

TABLE XIV. ESSD values and standard errors, not accounting for educational debts.

Factor	Item	Asian				Black				Hispanic				White Hispanic				White			
		Men		Women		Men		Women		Men		Women		Men		Women		Men		Women	
		ESSD	SE	ESSD	SE	ESSD	SE	ESSD	SE	ESSD	SE	ESSD	SE	ESSD	SE	ESSD	SE	ESSD	SE	ESSD	SE
First law	6	-0.64	0.11	-1.06	0.12	-0.96	0.21	-1.60	0.21	-1.05	0.19	-0.84	0.27	-0.27	0.15	-1.23	0.20	-0.66	0.09	-0.66	0.09
First law	7	-0.53	0.09	-0.69	0.15	-1.00	0.19	-1.35	0.23	-0.96	0.19	-0.78	0.23	-0.43	0.13	-0.72	0.22	-0.48	0.09	-0.48	0.09
First law	8	-0.49	0.09	-0.73	0.13	-0.90	0.19	-1.52	0.24	-1.00	0.19	-1.00	0.25	-0.32	0.13	-0.71	0.22	-0.53	0.09	-0.53	0.09
First law	10	-0.74	0.09	-0.89	0.13	-1.24	0.20	-1.42	0.22	-1.18	0.18	-1.02	0.24	-0.44	0.12	-1.50	0.20	-0.61	0.08	-0.61	0.08
First law	20	-0.14	0.09	-0.16	0.13	-0.65	0.17	-1.04	0.23	-0.66	0.18	-0.58	0.19	-0.40	0.10	-0.64	0.18	-0.04	0.07	-0.04	0.07
First law	23	-0.61	0.09	-1.12	0.15	-0.97	0.18	-1.93	0.29	-1.09	0.19	-1.07	0.24	-0.32	0.12	-1.21	0.20	-1.00	0.08	-1.00	0.08
First law	24	-0.59	0.09	-0.93	0.14	-1.11	0.21	-1.46	0.23	-0.82	0.18	-0.85	0.24	-0.28	0.13	-1.13	0.21	-0.52	0.08	-0.52	0.08
Second law	9	-0.11	0.08	0.00	0.00	-0.37	0.15	-0.22	0.18	-0.77	0.17	-0.20	0.16	-0.24	0.10	-0.07	0.18	-0.03	0.07	-0.03	0.07
Second law	12	-0.51	0.11	-1.02	0.14	-0.85	0.23	-1.39	0.21	-0.76	0.20	-1.13	0.23	-0.64	0.14	-1.39	0.19	-0.75	0.09	-0.75	0.09
Second law	14	-0.40	0.09	-1.16	0.14	-1.08	0.19	-1.72	0.27	-1.09	0.19	-1.15	0.25	-0.44	0.11	-1.14	0.23	-0.81	0.08	-0.81	0.08
Second law	19	-0.39	0.09	-0.60	0.13	-0.84	0.18	-0.97	0.22	-1.03	0.18	-0.67	0.19	-0.49	0.10	-0.66	0.20	-0.16	0.08	-0.16	0.08
Second law	21	-0.15	0.08	-0.29	0.12	-0.29	0.15	-0.61	0.18	-0.37	0.15	-0.25	0.16	-0.13	0.09	-0.19	0.17	-0.17	0.07	-0.17	0.07
Second law	22	-0.45	0.08	-0.66	0.13	-0.68	0.16	-0.97	0.21	-0.94	0.17	-0.52	0.22	-0.44	0.11	-0.98	0.22	-0.55	0.08	-0.55	0.08
Second law	27	-0.71	0.09	-1.15	0.14	-0.93	0.19	-2.12	0.31	-0.83	0.17	-1.54	0.29	-0.38	0.12	-1.24	0.23	-0.54	0.08	-0.54	0.08
Third law	4	-0.28	0.08	-0.27	0.12	-0.43	0.17	0.00	0.00	0.00	0.00	0.00	0.00	-0.14	0.11	-0.07	0.18	-0.03	0.07	-0.03	0.07
Third law	15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.22	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Third law	16	-0.48	0.09	-0.48	0.13	-0.68	0.19	-0.66	0.21	-0.50	0.18	-0.24	0.24	-0.21	0.13	-0.41	0.22	0.00	0.00	0.00	0.00
Third law	28	-0.31	0.08	-0.30	0.12	-0.57	0.16	-0.86	0.20	-0.57	0.15	-0.60	0.20	-0.41	0.10	-0.38	0.17	-0.19	0.06	-0.19	0.06
Identify forces	5	-0.23	0.08	-0.22	0.11	-0.72	0.15	-0.55	0.18	-0.87	0.16	-0.16	0.17	-0.26	0.09	-0.16	0.18	-0.09	0.07	-0.09	0.07
Identify forces	11	-0.34	0.08	-0.43	0.12	-0.60	0.17	-0.99	0.21	-0.48	0.16	-0.50	0.16	-0.35	0.10	-0.69	0.18	0.00	0.00	0.00	0.00
Identify forces	13	-0.48	0.08	-0.71	0.12	-0.99	0.18	-1.24	0.23	-1.01	0.16	-0.66	0.18	-0.49	0.10	-0.98	0.20	-0.30	0.06	-0.30	0.06
Identify forces	18	-0.13	0.07	-0.16	0.12	-0.75	0.15	-0.74	0.20	-0.72	0.16	-0.44	0.17	-0.38	0.10	-0.63	0.18	-0.13	0.07	-0.13	0.07
Identify forces	30	-0.16	0.08	-0.50	0.11	-0.30	0.15	-0.58	0.19	-0.59	0.16	-0.42	0.16	-0.28	0.09	-0.41	0.17	-0.06	0.07	-0.06	0.07
Superposition	17	0.00	0.00	0.11	0.10	0.00	0.00	-0.13	0.18	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.04	0.07	-0.04	0.07
Superposition	25	-0.08	0.07	0.00	0.00	-0.27	0.14	-0.86	0.18	-0.64	0.16	-0.37	0.17	-0.09	0.08	0.00	0.00	-0.13	0.07	-0.13	0.07
Superposition	26	0.00	0.00	-0.15	0.10	-0.40	0.14	-0.47	0.17	-0.65	0.14	0.00	0.00	-0.27	0.08	-0.27	0.14	-0.23	0.06	-0.23	0.06
None	1	-0.23	0.10	-0.46	0.15	-0.47	0.20	-0.79	0.21	-0.83	0.18	-0.52	0.24	-0.32	0.13	-0.53	0.23	-0.30	0.09	-0.30	0.09
None	2	-0.17	0.08	-0.29	0.13	0.00	0.00	0.00	0.00	-0.49	0.16	-0.15	0.17	0.00	0.00	-0.11	0.20	-0.33	0.08	-0.33	0.08
None	3	-0.58	0.09	-0.78	0.13	-0.84	0.17	-1.19	0.22	-0.91	0.18	-0.63	0.20	-0.51	0.12	-0.83	0.19	-0.40	0.08	-0.40	0.08
None	29	-0.11	0.10	0.02	0.16	0.24	0.23	0.02	0.24	0.20	0.20	0.27	0.28	0.20	0.14	0.34	0.22	0.26	0.09	0.26	0.09

TABLE XV. Infit and outfit statistics for the Rasch models, not accounting for educational debt.

Factor	Item	Asian				Black				Hispanic				White Hispanic				White			
		Men		Women		Men		Women		Men		Women		Men		Women		Men		Women	
		Infit	Outfit	Infit	Outfit	Infit	Outfit	Infit	Outfit	Infit	Outfit	Infit	Outfit	Infit	Outfit	Infit	Outfit	Infit	Outfit	Infit	Outfit
First law	6	0.99	0.91	0.86	0.79	0.93	0.99	1.00	1.02	0.94	0.78	0.98	0.74	0.88	0.64	0.92	0.76	0.97	0.72	0.93	0.81
First law	7	0.89	0.71	1.02	0.98	1.02	1.04	0.90	0.84	0.99	0.96	0.82	0.64	0.96	0.87	1.06	1.00	0.99	0.98	1.02	1.08
First law	8	0.97	0.87	0.97	0.95	0.99	0.94	0.95	0.89	0.98	0.95	0.95	0.82	0.95	0.84	1.17	1.18	1.00	0.89	1.05	1.01
First law	10	0.82	0.65	0.85	0.69	0.94	0.98	0.90	0.81	0.88	0.80	0.89	0.90	0.93	0.78	0.87	0.80	0.86	0.64	0.91	0.81
First law	20	0.96	0.93	0.95	0.87	0.92	0.88	1.05	1.05	0.99	0.99	0.95	0.87	0.91	0.85	0.98	1.04	0.95	0.92	0.97	1.01
First law	23	0.96	0.90	0.94	0.90	1.01	0.98	0.94	0.94	1.06	1.03	0.89	0.78	0.96	0.84	0.96	0.90	0.92	0.81	0.92	0.89
First law	24	0.91	0.86	0.96	0.86	0.96	0.98	0.98	1.04	0.94	1.02	0.99	1.06	0.94	0.97	0.91	0.83	0.84	0.61	0.96	0.84
Second law	9	0.87	0.81	1.09	1.12	0.92	0.86	0.86	0.80	1.03	1.06	0.85	0.78	0.92	0.86	0.97	0.91	0.96	0.91	0.98	0.94
Second law	12	0.98	0.96	0.93	0.80	1.11	1.01	0.99	1.03	0.98	1.08	1.04	1.15	0.96	0.92	0.98	1.05	1.00	1.02	1.02	1.01
Second law	14	1.01	0.97	1.04	1.01	1.01	1.03	0.94	0.93	0.95	0.92	1.19	1.31	1.05	1.09	0.92	0.89	1.06	1.04	1.02	0.99
Second law	19	0.92	0.89	1.00	1.03	0.83	0.77	0.97	1.01	0.89	0.84	1.05	1.05	0.87	0.79	1.00	0.97	0.89	0.76	0.91	0.87
Second law	21	0.99	1.02	0.93	0.97	0.96	1.00	0.82	0.77	0.85	0.86	1.04	1.08	0.96	0.95	1.02	1.15	1.07	1.11	1.03	1.07
Second law	22	1.02	0.96	1.05	1.04	1.05	1.04	0.94	0.88	0.97	0.92	1.23	1.13	0.95	0.85	0.98	1.04	0.98	0.85	1.02	1.00
Second law	27	1.01	0.95	0.97	0.98	1.07	1.06	1.08	1.18	0.98	0.91	0.82	0.78	1.01	1.00	1.10	1.14	1.00	0.94	1.07	1.03
Third law	4	0.84	0.89	0.86	0.77	0.89	0.82	1.07	1.16	1.01	1.07	0.91	1.36	0.93	0.84	0.84	0.84	0.95	0.88	0.91	0.80
Third law	15	1.10	1.09	1.22	1.19	0.97	0.94	0.99	0.97	1.06	1.08	1.16	1.07	1.21	1.26	1.10	1.18	1.22	1.25	1.10	1.12
Third law	16	0.93	0.94	0.93	0.86	0.92	0.96	1.05	1.09	0.91	0.82	0.97	1.29	0.98	1.13	0.89	0.72	0.95	0.91	0.95	0.86
Third law	28	0.81	0.73	0.80	0.68	0.76	0.68	0.97	0.89	0.84	0.80	0.80	0.66	0.82	0.69	0.82	0.82	0.79	0.63	0.84	0.73
Identify forces	5	0.99	1.02	0.92	0.91	0.84	0.84	0.98	0.93	0.94	1.01	0.88	0.82	0.92	0.90	0.95	0.93	0.92	0.88	0.89	0.85
Identify forces	11	0.93	0.90	0.93	0.91	0.89	0.89	0.93	0.93	0.87	0.86	0.93	0.83	1.01	1.03	0.91	0.83	0.92	0.90	0.85	0.81
Identify forces	13	0.82	0.77	0.73	0.67	0.86	0.83	0.85	0.81	0.77	0.74	0.81	0.78	0.77	0.69	0.78	0.70	0.74	0.63	0.74	0.66
Identify forces	18	0.94	0.88	0.85	0.77	0.80	0.71	0.95	0.94	0.93	0.88	0.70	0.64	0.81	0.73	0.84	0.78	0.84	0.74	0.78	0.70
Identify forces	30	1.04	1.03	0.87	0.87	0.86	0.87	1.02	1.05	0.95	0.96	0.86	0.77	0.82	0.79	0.86	0.86	0.89	0.87	0.91	0.89
Superposition	17	0.98	1.02	1.03	0.99	1.15	1.24	0.93	0.93	0.90	0.88	1.12	1.11	1.10	1.10	0.97	0.97	0.98	1.01	0.96	0.96
Superposition	25	0.97	0.94	1.09	1.12	1.05	1.14	0.87	0.86	0.95	0.98	1.19	1.48	0.93	0.92	0.86	0.82	0.88	0.87	0.90	0.90
Superposition	26	0.76	0.74	0.75	0.71	0.76	0.75	0.71	0.65	0.72	0.61	0.66	0.73	0.83	0.82	0.74	0.63	0.79	0.75	0.75	0.71
None	1	1.02	1.03	0.91	0.75	0.88	0.67	0.88	0.73	1.03	1.02	0.95	0.81	0.99	0.83	0.98	0.93	1.02	0.93	0.98	1.00
None	2	1.06	1.02	1.02	1.02	1.06	1.04	1.01	0.94	1.09	1.07	0.89	0.76	1.10	1.06	1.08	1.03	1.07	1.00	1.05	1.02
None	3	1.01	1.02	0.96	0.86	0.90	0.91	0.94	0.94	0.99	0.99	0.90	0.86	0.98	0.97	0.98	0.95	1.06	1.12	1.02	0.99
None	29	1.01	1.22	1.00	1.19	1.08	1.55	0.96	1.07	1.00	1.52	1.04	1.92	1.05	1.41	1.00	1.28	1.19	1.68	1.07	1.50

- [1] Jennifer L. Docktor and José P. Mestre, Synthesis of discipline-based education research in physics, *Phys. Rev. ST Phys. Educ. Res.* **10**, 020119 (2014).
- [2] Adrian Madsen, Sarah B. McKagan, and Eleanor C. Sayre, Resource letter RBAI-1: Research-based assessment instruments in physics and astronomy, *Am. J. Phys.* **85**, 245 (2017).
- [3] Eric Brewe, Vashti Sawtelle, Laird H. Kramer, George E. O'Brien, Idaykis Rodriguez, and Priscilla Pamelá, Toward equity through participation in modeling instruction in introductory university physics, *Phys. Rev. ST Phys. Educ. Res.* **6**, 010106 (2010).
- [4] Adrienne Traxler and Eric Brewe, Equity investigation of attitudinal shifts in introductory physics, *Phys. Rev. ST Phys. Educ. Res.* **11**, 020132 (2015).
- [5] David J. Webb and Cassandra A. Paul, Attributing equity gaps to course structure in introductory physics, *Phys. Rev. Phys. Educ. Res.* **19**, 020126 (2023).
- [6] Ben Van Dusen, Jayson Nissen, Robert M. Talbot, Hannah Huvard, and Mollee Shultz, A quantcrit investigation of society's educational debts due to racism and sexism in chemistry student learning, *J. Chem. Educ.* **99**, 25 (2021).
- [7] Ben Van Dusen and Jayson Nissen, Equity in college physics student learning: A critical quantitative intersectionality investigation, *J. Res. Sci. Teach.* **57**, 33 (2020).
- [8] Ben Van Dusen and Jayson Nissen, How statistical model development can obscure inequities in stem student outcomes, *J. Women Minorities Sci. Eng.* **28**, 27 (2022).
- [9] B. Van Dusen and J. Nissen, Associations between learning assistants, passing introductory physics, and equity: A quantitative critical race theory investigation, *Phys. Rev. Phys. Educ. Res.* **16**, 010117 (2020).
- [10] Jayson M. Nissen, Ian Her Many Horses, and Ben Van Dusen, Investigating society's educational debts due to racism and sexism in student attitudes about physics using quantitative critical race theory, *Phys. Rev. Phys. Educ. Res.* **17**, 010116 (2021).
- [11] Stephen Kanim and Ximena C. Cid, Demographics of physics education research, *Phys. Rev. Phys. Educ. Res.* **16**, 020106 (2020).
- [12] A. Traxler, R. Henderson, J. Stewart, G. Stewart, A. Papak, and R. Lindell, Gender fairness within the force concept inventory, *Phys. Rev. Phys. Educ. Res.* **14**, 010103 (2018).
- [13] Rachel Henderson, Paul Miller, John Stewart, Adrienne Traxler, and Rebecca Lindell, Item-level gender fairness in the force and motion conceptual evaluation and the conceptual survey of electricity and magnetism, *Phys. Rev. Phys. Educ. Res.* **14**, 020103 (2018).
- [14] Philip Eaton, Evidence of measurement invariance across gender for the force concept inventory, *Phys. Rev. Phys. Educ. Res.* **17**, 010130 (2021).
- [15] Nicole M. Else-Quest and Janet Shibley Hyde, Intersectionality in quantitative psychological research: II. Methods and techniques, *Psychol. Women Q.* **40**, 319 (2016).
- [16] Patricia Hill Collins and Sirma Bilge, *Intersectionality* (John Wiley & Sons, New York, 2020).
- [17] David Hestenes, Malcolm Wells, and Gregg Swackhamer, Force concept inventory, *Phys. Teach.* **30**, 141 (1992).
- [18] Philip Eaton and Shannon D. Willoughby, Confirmatory factor analysis applied to the force concept inventory, *Phys. Rev. Phys. Educ. Res.* **14**, 010124 (2018).
- [19] Alicen Morley, Jayson M. Nissen, and Ben Van Dusen, Measurement invariance across race and gender for the force concept inventory, *Phys. Rev. Phys. Educ. Res.* **19**, 020102 (2023).
- [20] D. Andrich and C. Hagquist, Real and artificial differential item functioning, *J. Educ. Behav. Stud.* **37**, 387 (2012).
- [21] Diane L. Putnick and Marc H. Bornstein, Measurement invariance conventions and reporting: The state of the art and future directions for psychological research, *Dev. Rev.* **41**, 71 (2016).
- [22] Yadolah Dodge, David Cox, and Daniel Commenges, *The Oxford Dictionary of Statistical Terms* (Oxford University Press on Demand, 2003).
- [23] Gloria Ladson-Billings, From the achievement gap to the education debt: Understanding achievement in US schools, *Educ. Res.* **35**, 3 (2006).
- [24] Gloria Ladson-Billings, Pushing past the achievement gap: An essay on the language of deficit, *J. Negro Educ.* **76**, 316 (2007).
- [25] Kimberle Crenshaw, Mapping the margins: Intersectionality, identity politics, and violence against women of color, *Stan. L. Rev.* **43**, 1241 (1990).
- [26] Leslie McCall, The complexity of intersectionality, *Signs* **30**, 1771 (2005).
- [27] Clare R. Evans, David R. Williams, Jukka-Pekka Onnela, and S. V. Subramanian, A multilevel approach to modeling health inequalities at the intersection of multiple social identities, *Soc. Sci. Med.* **203**, 64 (2018).
- [28] Jennifer C. Nash, Re-thinking intersectionality, *FEMS Microbiol. Rev.* **89**, 1 (2008).
- [29] Maria Ong, *The Double Bind in Physics Education: Intersectionality, Equity, and Belonging for Women of Color* (Harvard Education Press, 2023).
- [30] Ebony Omotola McGee, *Black, Brown, Bruised: How Racialized STEM Education Stifles Innovation* (Harvard Education Press, 2021).
- [31] American Educational Research Association, American Psychological Association, National Council on Measurement in Education, and Joint Committee on Standards for Educational and Psychological Testing (U.S.), *Standards for Educational and Psychological Testing* (American Educational Research Association, 2014).
- [32] William H. Angoff, *Validity: An Evolving Concept*, 1st ed. (Routledge, 2013), pp. 19–32.
- [33] G. Camilli, Test fairness, in *Educational Measurement*, edited by L. R. Brennan (American Council on Education, Westport, CT, 2006), pp. 20–256.
- [34] Bruno D. Zumbo, Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going, *Lang. Assess. Q.* **4**, 223 (2007).
- [35] Paul W. Holland, Causation and race, *ETS Res. Rep. Ser.* **2003**, i–21 (2003).
- [36] P. H. McCalliser, Testing, dif, and public policy, in *Differential Item Functioning*, edited by Paul W. Holland and Howard Wainer (Earlbaum Associates, Hillsdale, NJ, 1993), pp. 389–396.

- [37] Nathan Mantel and William Haenszel, Statistical aspects of the analysis of data from retrospective studies of disease, *J. Natl. Cancer Inst.* (1940-1978) **22**, 719 (1959).
- [38] John Brodersen, David Meads, Svend Kreiner, Hanne Thorsen, Lynda Doward, and Stephen McKenna, Methodological aspects of differential item functioning in the Rasch model, *J. Med. Econ* **10**, 309 (2007).
- [39] Sharon E. Osborn Popp, David E. Meltzer, and Colleen Megowan-Romanowicz, Is the force concept inventory biased? Investigating differential item functioning on a test of conceptual learning in physics documents, in *American Educational Research Association Conference* (2011), https://physicseducation.net/talks/FCI_DIF_AERA_2011.pdf.
- [40] Rachel Henderson, John Stewart, and Adrienne Traxler, Partitioning the gender gap in physics conceptual inventories: Force concept inventory, force and motion conceptual evaluation, and conceptual survey of electricity and magnetism, *Phys. Rev. Phys. Educ. Res.* **15**, 010131 (2019).
- [41] Jennifer Docktor and Kenneth Heller, Gender differences in both force concept inventory and introductory physics performance, *AIP Conf. Proc.* **1064**, 15 (2008).
- [42] Physport assessments: Force concept inventory, <https://www.physport.org/assessments/assessment.cfm?A=FCI>.
- [43] LASSO (2025), <https://www.lassoeducation.org>.
- [44] Richard R. Hake, Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses, *Am. J. Phys.* **66**, 64 (1998).
- [45] Yang Xiao, Guiqing Xu, Jing Han, Hua Xiao, Jianwen Xiong, and Lei Bao, Assessing the longitudinal measurement invariance of the force concept inventory and the conceptual survey of electricity and magnetism, *Phys. Rev. Phys. Educ. Res.* **16**, 020103 (2020).
- [46] Jesper Bruun and Eric Brewwe, Talking and learning physics: Predicting future grades from network measures and force concept inventory pretest scores, *Phys. Rev. ST Phys. Educ. Res.* **9**, 020109 (2013).
- [47] Marcos D. Caballero, Edwin F. Greco, Eric R. Murray, Keith R. Bujak, M. Jackson Marr, Richard Catrambone, Matthew A. Kohlmyer, and Michael F. Schatz, Comparing large lecture mechanics curricula using the force concept inventory: A five thousand student study, *Am. J. Phys.* **80**, 638 (2012).
- [48] Jayson M. Nissen, Ian Her Many Horses, Ben Van Dusen, Manher Jariwala, and Eleanor Close, Providing context for identifying effective introductory mechanics courses, *Phys. Teach.* **60**, 179 (2022).
- [49] Melanie Good, Alexandru Maries, and Chandralekha Singh, Impact of traditional or evidence-based active-engagement instruction on introductory female and male students' attitudes and approaches to physics problem solving, *Phys. Rev. Phys. Educ. Res.* **15**, 020129 (2019).
- [50] Eric Brewwe, Laird Kramer, and George O'Brien, Modeling instruction: Positive attitudinal shifts in introductory physics measured with class, *Phys. Rev. ST Phys. Educ. Res.* **5**, 013102 (2009).
- [51] Philip Eaton and Shannon D. Willoughby, Confirmatory factor analysis applied to the force concept inventory, *Phys. Rev. Phys. Educ. Res.* **14**, 010124 (2018).
- [52] Philip Eaton and Shannon Willoughby, Identifying a preinstruction to postinstruction factor model for the force concept inventory within a multitrait item response theory framework, *Phys. Rev. Phys. Educ. Res.* **16**, 010106 (2020).
- [53] Jing Wang and Lei Bao, Analyzing force concept inventory with item response theory, *Am. J. Phys.* **78**, 1064 (2010).
- [54] Maja Planinic, Lana Ivanjek, and Ana Susac, Rasch model based analysis of the force concept inventory, *Phys. Rev. ST Phys. Educ. Res.* **6**, 010103 (2010).
- [55] Ben Van Dusen, LASSO: A new tool to support instructors and researchers, [arXiv:1812.02299](https://arxiv.org/abs/1812.02299).
- [56] Sasha Shen Johfre and Jeremy Freese, Reconsidering the reference category, *Sociol. Methodol.* **51**, 253 (2021).
- [57] Neil J. Dorans and Paul W. Holland, DIF detection and description: Mantel Haenszel and standardization, Educational Testing Service, Princeton, NJ, Technical Report No. RR-92-10, 1992, <https://files.eric.ed.gov/fulltext/ED387526.pdf>.
- [58] Rebecca Zwick, A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement, Educational Testing Service, Princeton, NJ, Technical Report No. RR-12-08, 2012, <https://files.eric.ed.gov/fulltext/EJ1109842.pdf>.
- [59] D. Magis, S. Beland, F. Tuerlinckx, and P. De Boeck, A general framework and an R package for the detection of dichotomous differential item functioning, *Behav. Res. Meth. Instrum. Comput.* **42**, 847 (2010).
- [60] Tam H. Nguyen, Hae-Ra Han, Miyong T. Kim, and Kitty S. Chan, An introduction to item response theory for patient-reported outcome measurement, *Patient Educ. Couns.* **7**, 23 (2014).
- [61] J. Kopf, A. Zeileis, and C. Strobl, Anchor selection strategies for DIF analysis: Review, assessment, and new approaches, *Educ. Psychol. Meas.* **75**, 22 (2015).
- [62] C. Strobl, J. Kopf, L. Kohler, T. von Oertzen, and A. Zeileis, Anchor point selection: Scale alignment based on an inequality criterion, *Appl. Psychol. Meas.* **45**, 214 (2021).
- [63] Idaykis Rodriguez, Eric Brewwe, Vashti Sawtelle, and Laird H. Kramer, Impact of equity models and statistical measures on interpretations of educational reform, *Phys. Rev. ST Phys. Educ. Res.* **8**, 020103 (2012).
- [64] Michael J. Cahill, K. Mairin Hynes, Rebecca Trousil, Lisa A. Brooks, Mark A. McDaniel, Michelle Repice, Jiuqing Zhao, and Regina F. Frey, Multiyear, multi-instructor evaluation of a large-class interactive-engagement curriculum, *Phys. Rev. ST Phys. Educ. Res.* **10**, 020101 (2014).
- [65] Laura McCullough, Gender differences in student responses to physics conceptual questions based on question context, in *Proceedings of the ASQ Advancing the STEM Agenda in Education, the Workplace and Society, University of Wisconsin-Stout* (2011), https://www.academia.edu/857468/Gender_Differences_in_Student_Responses_to_Physics_Conceptual_Questions_Based_on_Question_Context.
- [66] Shanda Lauer, Jennifer Momsen, Erika Offerdahl, Mila Kryjevskaja, Warren Christensen, and Lisa Montplaisir, Stereotyped: Investigating gender in introductory science courses, *CBE-Life Sci. Educ.* **12**, 30 (2013).

- [67] Simon Bates, Robyn Donnelly, Cait MacPhee, David Sands, Marion Birch, and Niels R. Walet, Gender differences in conceptual understanding of Newtonian mechanics: A UK cross-institution comparison, *Eur. J. Phys.* **34**, 421 (2013).
- [68] Alma Robinson, John H. Simonetti, Kasey Richardson, and Megan Wawro, Positive attitudinal shifts and a narrowing gender gap: Do expertlike attitudes correlate to higher learning gains for women in the physics classroom?, *Phys. Rev. Phys. Educ. Res.* **17**, 010101 (2021).
- [69] John M. Linacre *et al.*, What do infit and outfit, mean-square and standardized mean, Rasch Meas. Trans. **16**, 878 (2002).
- [70] Yves Rosseel, Daniel Oberski, Jarrett Byrnes, Leonard Vanbrabant, Victoria Savalei, Ed Merkle, Michael Hallquist, Mijke Rhemtulla, Myrsini Katsikatsou, and Barendse Mariska *et al.*, Package 'lavaan', Retrieved June 17, 2017.
- [71] Wen-Hung Chen and David Thissen, Local dependence indexes for item pairs using item response theory, *J. Educ. Behav. Stat.* **22**, 265 (1997).
- [72] Philip Eaton, Barrett Frank, and Shannon Willoughby, Detecting the influence of item chaining on student responses to the force concept inventory and the force and motion conceptual evaluation, *Phys. Rev. Phys. Educ. Res.* **16**, 020122 (2020).
- [73] Jacob Cohen, *Statistical Power Analysis for the Behavioral Sciences* (Routledge, 2013).
- [74] Adam W. Meade, A taxonomy of effect size measures for the differential functioning of items and scales, *J. Appl. Psych.* **95**, 728 (2010).
- [75] R. Philip Chalmers, mirt: A multidimensional item response theory package for the R environment, *J. Stat. Softw.* **48**, 1 (2012).
- [76] Peter H. Westfall, Randall D. Tobias, and Russell D. Wolfinger, *Multiple Comparisons and Multiple Tests Using SAS* (SAS Institute, 2011).
- [77] R. L. Wasserstein and N. A. Lazar *et al.*, The ASA's statement on p -values: Context, process, and purpose, *Am. Stat.* **70**, 129 (2016).
- [78] Valentin Amrhein, David Trafimow, and Sander Greenland, Inferential statistics as descriptive statistics: There is no replication crisis if we don't expect replication, *Am. Stat.* **73**, 262 (2019).
- [79] R. L. Wasserstein, A. L. Schirm, and N. A. Lazar, Moving to a world beyond $p < 0.05$, *Am. Stat.* **73**, 1 (2019).
- [80] Sander Greenland, Valid p -values behave exactly as they should: Some misleading criticisms of p -values and their resolution with s -values, *Am. Stat.* **73**, 106 (2019).
- [81] Ronald K. Thornton and David 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).
- [82] Ben Van Dusen and Jayson Nissen, Equity in college physics student learning: A critical quantitative intersectionality investigation, *J. Res. Sci. Teach.* **57**, 33 (2020).
- [83] Charles Henderson and Melissa H. Dancy, Impact of physics education research on the teaching of introductory quantitative physics in the united states, *Phys. Rev. ST Phys. Educ. Res.* **5**, 020107 (2009).
- [84] Ronald K. Thornton, Dennis Kuhl, Karen Cummings, and Jeffrey Marx, Comparing the force and motion conceptual evaluation and the force concept inventory, *Phys. Rev. ST Phys. Educ. Res.* **5**, 010105 (2009).
- [85] Adrian Madsen, Sarah B. McKagan, and Eleanor C. Sayre, Gender gap on concept inventories in physics: What is consistent, what is inconsistent, and what factors influence the gap?, *Phys. Rev. ST Phys. Educ. Res.* **9**, 020121 (2013).
- [86] Douglas Huffman and Patricia Heller, What does the force concept inventory actually measure?, *Phys. Teach.* **33**, 138 (1995).
- [87] Robert Krakehl and Angela M. Kelly, Intersectional analysis of advanced placement physics participation and performance by gender and ethnicity, *Phys. Rev. Phys. Educ. Res.* **17**, 020105 (2021).
- [88] E. Seymour and N. M. Hewitt, *Talking About Leaving: Why Undergraduates Leave the Sciences*, 1st ed. (Westview Press, Boulder, CO, 1997), ISBN-10 0813389267, ISBN-13 978-0813389264.
- [89] David J. Webb, Cassandra A. Paul, and Mary K. Chessey, Relative impacts of different grade-scales on student success in introductory physics, [arXiv:1903.06747](https://arxiv.org/abs/1903.06747).
- [90] Cassandra A. Paul and David J. Webb, Percent grade scale amplifies racial or ethnic inequities in introductory physics, *Phys. Rev. Phys. Educ. Res.* **18**, 020103 (2022).
- [91] Akira Miyake, Lauren E. Kost-Smith, Noah D. Finkelstein, Steven J. Pollock, Geoffrey L. Cohen, and Tiffany A. Ito, Reducing the gender achievement gap in college science: A classroom study of values affirmation, *Science* **330**, 1234 (2010).
- [92] Danny Doucette and Chandralekha Singh, Share it, don't split it: Can equitable group work improve student outcomes?, *Phys. Teach.* **60**, 166 (2022).
- [93] Lisabeth M. Santana and Chandralekha Singh, Effects of male-dominated physics culture on undergraduate women, *Phys. Rev. Phys. Educ. Res.* **19**, 020134 (2023).
- [94] Carlos Fernando Collares, Cognitive diagnostic modelling in healthcare professions education: An eye-opener, *Adv. Health Sci. Educ.* **27**, 427 (2022).
- [95] Robert J. Mislevy, Russell G. Almond, and Janice F. Lukas, A brief introduction to evidence-centered design, *ETS Res. Rep. Ser.* **2003**, i–29 (2003).
- [96] Benjamin Pollard, Robert Hobbs, Rachel Henderson, Marcos D. Caballero, and H. J. Lewandowski, Introductory physics lab instructors' perspectives on measurement uncertainty, *Phys. Rev. Phys. Educ. Res.* **17**, 010133 (2021).
- [97] Michael Vignal, Gayle Geschwind, Benjamin Pollard, Rachel Henderson, Marcos D. Caballero, and H. J. Lewandowski, Survey of physics reasoning on uncertainty concepts in experiments: An assessment of measurement uncertainty for introductory physics labs, *Phys. Rev. Phys. Educ. Res.* **19**, 020139 (2023).