Measurement invariance across race and gender for the Force Concept Inventory

Alicen Morley,[1] Jayson M. Nissen,[2] and Ben Van Dusen[1]

[1]*School of Education, Iowa State University, Ames, Iowa 50011, USA*
[2]*Nissen Education and Research Design, Slidell, Louisiana 70458, USA*

(Received 5 November 2022; accepted 8 June 2023; published 6 July 2023)

Instructors and researchers often use research-based assessments to identify the impact of instructional activities. These investigations often focus on issues of diversity, equity, and inclusions by comparing outcomes across social identity groups (e.g., gender, race, and class). Comparisons across groups assume the assessments measure the same factors in the same way across social identity groups. Very few research-based assessments, however, have validation evidence to support this assumption. Measurement invariance testing provides validity evidence that an assessment measures latent factors (e.g., Newtonian thinking or physics identity) equivalently across groups. We examined the measurement invariance of the Eaton and Willoughby five (EW5) factor model on the Force Concept Inventory. We found evidence for measurement invariance across the intersections of gender (men and women) and race or ethnicity (Asian, Black, Hispanic, White, and White Hispanic). These results indicate that performance differences across the five factors can further understanding of equity in physics courses. Without measurement invariance, such work could produce misleading results that undermine efforts to support equity in physics courses.

DOI: 10.1103/PhysRevPhysEducRes.19.020102

I. INTRODUCTION

Research-based assessments (RBAs) have provided the foundation for many physics course pedagogical and curricular transformations [1,2]. The primary feature that distinguishes a RBA from a typical assessment is its validation arguments [3]. Researchers use RBAs to collect data for investigations of diversity, equity, and inclusion, e.g., [4–6]. Instrument developers, however, rarely examine instrument validity across social identity groups, such as gender and race [7,8]. Most physics RBAs lack validation arguments for many groups, including Black, Hispanic, Indigenous, and nonbinary students, limited validation arguments for women, and no validation arguments across the intersection of race and gender, e.g., for White Hispanic Women. These limited validation arguments may hide biases on the assessments that create misleading findings about group outcomes.

Research on equity in physics tends to focus on the overall assessment scores, see Refs. [4–6,8,9]. Using overall scores, however, prevents studies from identifying the role of more specific latent factors in promoting equity of outcomes. Some studies, however, have investigated equity using the latent factors that an RBA measures. Latent factors are the underlying constructs in a student's mind (e.g., Newton's third law, conservation of momentum, or physics identity) that a RBA measures. Research on gender and physics identity often uses one instrument to look at factors for interest, performance, recognition, or competence, for example [10,11]. These factors provide a better understanding of inequities and systems or interventions that create or address those inequities. Research on conceptual understanding also uses subsets of questions from concept inventories to assess student learning, for example, Refs. [12,13]. Equity research on conceptual learning in physics, however, seldom uses the latent factors measured by RBAs, such as the Force Concept Inventory (FCI) [14].

Research on equity in physics sometimes considers the intersection of race and gender (e.g., [15–19]). Many physics equity research investigations, however, have aggregated students across many social identity groups [19] and have not examined the interactions between multiple axes of social identities. For example, when investigating inequities across race, many studies have compared the aggregated outcomes of White or White and Asian students against those of all other races [9,20–23]. While these studies can provide important insights, intersectionality theory [24,25] argues that these aggregations can obscure inequities and injustices. For example, Shafer *et al.* [26] found that the common approach of comparing underrepresented (URM) and non-URM groups hides the inequities that both Black and Asian students face in physics. Disaggregation across multiple intersecting axes of students' social identities (e.g., race and gender) and different power structures within physics (e.g., the focus of research on calculus-based physics courses and lack of research at two-year colleges) can support targeting and assessing interventions to create more equitable and just physics courses.

Identifying differences in factor outcomes across the intersections of race and gender could provide researchers and instructors with insights into ways to support students from multiply marginalized groups. Take, for example, a class where most of the students took a high school physics course, but a minority of students did not take such a course due to racism or class oppression through the underfunding of their schools. If the instructor administers the FCI and looks at the class average pretest score on Newton's laws (a topic covered extensively in high school physics courses), they may conclude that they only need to briefly review the topic before moving on to content that builds on the ideas. This data-driven decision would optimize the course for the "average" student but exacerbate inequities. If, instead, the instructor disaggregated the findings across social identity groups, they would see that spending more time on Newton's laws would help create a more equitable classroom that addresses the educational debts society owes to students [27]. Before instructors can do this, however, researchers must conduct measurement invariance testing and differential item function analysis [8,28] to establish that the factors on an instrument measure the same constructs across groups.

In this investigation, we use measurement invariance testing and reliability analysis to examine the validity of the FCI [14] across the intersections of race, gender, and calculus and algebra-based physics courses. Test validity covers a wide range of arguments [29]. We focus on two validation arguments central to RBAs: internal structure validity and reliability. Internal structure validity focuses on the accuracy of a measurement; does an assessment measure what researchers designed it to measure or is it measuring something else or several things (multidimensional). Reliability arguments focus on the precision of a measurement, the signal-to-noise ratio [30]. Measurement invariance analysis addresses the internal structure validity by informing whether an instrument measures factors consistently across groups or across time, e.g., pretest and post-test. Reliability analysis quantifies the precision of each factor [30]. Researchers often use Cronbach's alpha in their reliability analysis. Cronbach's alpha is the ratio of the total score variance explained by the factor divided by the estimated variance of the observed total score. The findings extend the internal structure validation arguments for the Eaton and Willoughby five (EW5) factor model [31] on the FCI across the intersection of race and gender while identifying limitations in the precision of those factors when groups have low mean test scores, such as for pretests.

II. THEORETICAL FRAMEWORK

Intersectionality grew out of Black Feminist theory and posits that understanding marginalized women's experiences requires viewing them through the lens of multiple systems of oppression, among them racism, sexism, heteronormativity, and class oppression [24,25]. For example, an intersectional analysis looks at the outcomes and experiences of Black lesbians as distinct from Black students, women, and LGBTQ + students [32]. Rocabado *et al.* [33] point to the need for equity research to account for intersections of multiple identities and power structures and to test for measurement invariance of the quantitative tools used in that research. We are, however, only aware of one measurement invariance study that explicitly drew on intersectionality theory [34,35].

Collins [36] argues that intersectionality is a theory in flux as it crosses disciplinary and methodological boundaries. Collins also, however, provides a provisional list of guiding assumptions that intersectionality studies embrace. Drawing on that list, this work assumes that race, gender, and class are best understood in relational terms, rather than in isolation from one another and that these categories are mutually constructed. Statistical power and minimum sample sizes limit how many systems of oppression different quantitative methods can investigate. Intersectionality's commitment to challenging the status quo of the inequalities created by these interacting systems of power, however, motivated our work to be as intersectional as possible within the constraints of our data and methods. In this goal, we align with the framework for intersectionality research laid out by Cho *et al.* [37] in that we are applying the lessons of intersectionality to existing common practices within physics education research: the use and validation of research-based assessments.

III. DEFINITIONS

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

IV. RESEARCH QUESTIONS

To determine if the FCI can identify differences across groups for both the pretest and post-test on the five latent factors that it measures, we asked the following questions:

- To what extent is the EW5 factor model on the FCI measurement invariant across the intersections of gender and race?
- To what extent are the five latent factors reliable across the intersections of gender and race?

If the results do not indicate measurement invariance and reliability for the EW5 model on the FCI, then instructors and researchers should not compare scores on the instrument factors across groups. The lack of invariance does not mean the entire instrument is problematic, though it does

TABLE I. Definitions of statistical and equity terms.

Term	Definition
Measurement invariance	Measurement invariance testing assesses the psychometric equivalence of a latent factor across groups or time [38]. In this case, we assess whether an instrument measures latent factors consistently across social identity groups.
Latent factor	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) [39].
Social identity group	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.
Structural equation model (SEM)	A set of statistical techniques that allow modeling and testing the relationships between observable and theoretical variables [40].
Confirmatory factor analysis (CFA)	A special case of SEM that examines the relationships between observed measures (e.g., item scores) and latent factors [41].
Fit indices	A quantitative measure of how well data fit a model. Examples from this paper include CFI, TLI, and RMSEA.
Validity	The extent to which evidence and theory support the interpretation of test scores for the proposed uses of the test [42].
Reliability	How precisely a combination of questions measures a latent factor for a population that is quantified as a ratio of the signal (content knowledge) to noise (error, e.g., guessing). [30,33].
Research-based assessment (RBA)	An assessment that has undergone rigorous development and testing with reliability and validity arguments [3].

warrant further studies such as those by Traxler *et al.* [8] to investigate the validity and reliability of the FCI. Measurement invariance and reliability of the EW5 model on the FCI would allow additional information from the factors to inform issues of equity in physics courses.

V. THE FORCE CONCEPT INVENTORY

Hestenes *et al.* [14] developed the FCI to probe student understanding of Newtonian forces to assess the effectiveness of physics instruction. Researchers have applied many different quantitative methods to data from the FCI, see Ref. [43]. One strand of this quantitative research focuses on the fairness of the FCI and its ability to give unbiased data across different groups of students. Evidence indicates that several items on the FCI function differently for men [8] and White men in particular [28] than for other social identity groups. Nonetheless, researchers have often used the FCI to investigate the effectiveness of instructional techniques [44–48] and equity in courses [4,49,50].

Hestenes *et al.* [14] proposed that the FCI covered six areas of Newtonian forces. Findings from multiple studies, however, suggest that it examines five factors [31,43,51,52]. Eaton and Willoughby [31] found that their five-factor model (EW5) provided the best fit for the data. The original EW5 model, shown in Table II, omitted questions 1, 2, 3, and 29 due to very poor fit and little explanatory value as shown by a number of studies [31,52]. Based on evidence from Xiao *et al.* [45], Eaton [51] updated their model to include those four questions and found factor validity was still acceptable with these four items included. Eaton refers to this model as the Eaton and Willoughby five factor modified (EW5M) model.

Wang and Bao [53] applied a three-parameter item response model to FCI pretests for introductory mechanics students. The three parameters informed how difficult the questions were, how well they discriminated between different ability levels, and how frequently a question was answered correctly due to guessing. Applying their results to the EW5 model indicates that the questions on factors 1 and 2 of the EW5 model tend to have easier questions with lower discrimination and higher likelihood of guessing the right answer than the questions on the other three factors. Item difficulties from Planinic *et al.* [54] also match this trend. These results indicate that the questions on factors 1 and 2 may provide poor data for students with lower scores. For example, Wang and Bao's results indicate

TABLE II. Eaton and Willoughby [31] 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

that 7 of the 14 items on factors 1 and 2 had guessing rates of 20% or higher. If a group of students had an average score of 40% on these items, then approximately half of their correct answers likely came from guessing. The guessing introduces noise and this noise consumes small signals. This inference aligns with Planinic *et al.* [54] finding that the FCI may function differently for populations of students with lower test scores.

The poor psychometrics of several items on the FCI and many easier items with high guessing rates loading on two of the factors indicate that the FCI has many limitations for collecting data from students with a wide range of abilities. The number of studies investigating the FCI, see Ref. [43], and the frequency that it is used, however, indicates that many instructors and researchers could use the additional data from the factors on the FCI to support equity in introductory physics courses. Measurement invariance on the FCI, in addition to the extensive psychometric analysis of the FCI, can also inform the design and development of an RBA that addresses the limitations of the FCI.

VI. MATERIALS AND METHODS

A. Data

We collected the data from the Learning About STEM Student Outcomes (LASSO) platform's anonymous research database [55,56]. 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 133 algebra-based and 194 calculus-based first-semester college physics courses across 47 institutions. The institutions were distributed across the four census tracks in the Northeast (13), South (10), Midwest (7), West (16), and outside the United States (1). We used the Carnegie Classification of Institutions of Higher Education (CCIHE) public 2021 database to characterize the 45 institutions in the database. For the two institutions not in the CCIHE, one was outside the United States and one was a branch campus. The dataset included data from 19 high or very high research-intensive institutions, 11 minority-serving institutions that were all Hispanic-serving, and 5 primarily associate degree-granting institutions. Of the 47 institutions, 34 were public and 13 were private not-for-profit institutions with 3 of these also being research intensive. The data included 25 large, 16 medium, and 6 small institutions.

LASSO also provided students' self-identified social identity information. Specifically, it included student responses to gender, race, ethnicity, and first or continuing generation college status questions. Students were allowed to select multiple options for gender and race and could write in their own answer for each question. As shown in Table III, approximately one in three students who identified as Hispanic left the race question blank. This led us to treat Hispanic similarly similar to the responses to the race question; students could identify as only Hispanic. This decision aligns with two-thirds of Hispanic adults considering Hispanic as part of their racial identity [57].

TABLE III. Sample sizes by race, gender, and course type.

Race	Math	Men	Women
Asian	Algebra	404	599
	Calculus	500	300
Black	Algebra	104	221
	Calculus	177	144
Hispanic	Algebra	111	142
	Calculus	268	90
White	Algebra	2949	3225
	Calculus	3251	1102
White Hispanic	Algebra	226	246
	Calculus	453	177

B. Data preparation

We removed scores for students who took less than 5 min on that assessment and students who did not complete either the pretest or post-test; the Appendix details the proportions of removed scores. Given a recommended minimum sample size of 200 participants [58], we were able to evaluate measurement invariance across five races (Asian, Black, Hispanic, White, and White Hispanic) and two genders (women and men). Table III details the sample sizes for the groups included in the analysis. We could not include American Indian, Middle Eastern, Pacific Islander, genderqueer, and nonbinary students nor could we differentiate between transgender and cisgender men and women. In measurement invariance testing, each student has to belong to a distinct group. While we had enough data from Hispanic, White Hispanic, and White students to distinguish these three groups, we did not have enough data to include other social identity groups, such as Black Hispanic students.

The data also included student responses for first or continuing-generation college status and if the physics course was algebra or calculus based. If we disaggregated groups further by first or continuing-generation status or course type, the dataset only met our minimum sample size for Asian and White men and women. Our preliminary results from measurement invariance testing across course types and first or continuing-generation status were similar to those presented in the results. We excluded this analysis from the manuscript for brevity.

C. Structural equation modeling with ordered categorical data

We used a structural equation modeling (SEM) approach to measurement invariance testing. For ordered categorical data, such as the right or wrong coding of the FCI data we used, or for Likert-scale data, SEM computes a polychoric

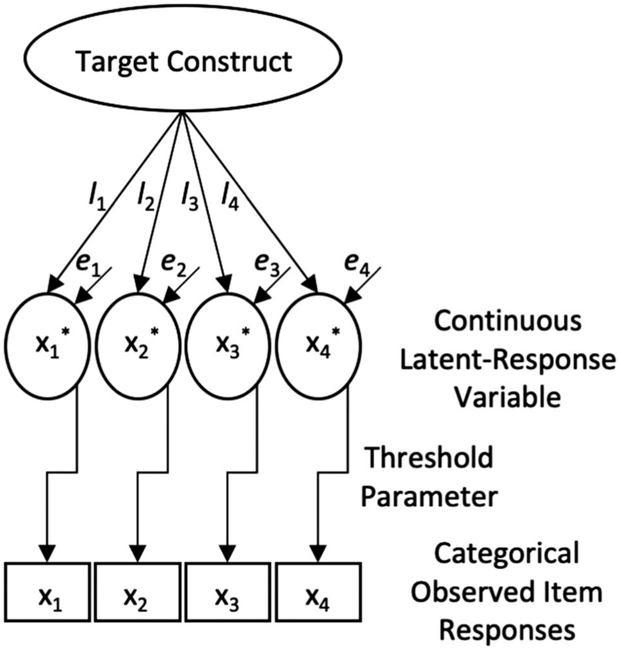

FIG. 1. Path diagram for a one-factor model consisting of four ordered categorical items. Path diagrams visualize the equations in SEM. The rectangles represent observed data. The ovals represent latent factors that were not directly measured. In SEM for ordered categorical data, the observed data are analyzed using continuous latent variables. The threshold parameters, τ, link the observed data to the continuous latent variable. Figure 2 presents two examples of thresholds and continuous latent-response variables. The straight arrows represent the factor loadings linking the continuous latent response variables to the target factor. The short diagonal arrow (e) pointing into the latent response variable is the residual variance.

correlation between each pair of ordinal variables [59]. This polychoric correlation assumes that a normally distributed, latent continuous variable underlies the observed ordinal data, as shown in Figs. 1 and 2. Figure 1 illustrates this assumption where the rectangles represent the observed ordered categorical responses and the ovals represent the underlying continuous latent variables. The observed responses and underlying latent variables are connected by a bent arrow to represent the threshold parameters (τ). Figure 2 illustrates these thresholds for both a binary item and a 6-point Likert-scale item. The thresholds (τ) divide the continuous latent variable into the response spaces. The values on that latent continuous variable then manifest as those ordered categories when they are measured. Standard practice sets the mean at 0 and the variance at 1 for latent continuous variables.

D. Measurement invariance

The structural equation modeling (SEM) approach to measurement invariance testing included four steps and three levels of invariance. Figure 3 provides examples of steps 2–4. In the first step, we ran confirmatory factor analysis (CFA) for each group to ensure the model fit the data for each group. In the second step, we fixed the configuration of which factors loaded on which items across all of the groups. Satisfying this step indicated the *configural invariance* of the model's overall structure across all of the groups. The third step fixed the value of the factor loadings across all of the groups. Satisfying this step established the *metric invariance* of the model's structure across groups. Because we modeled our data as categorical using placeholder variables of 0 for the wrong answer and 1 for the right answer, our final step fixed the thresholds for each item across all groups. This step established *scalar invariance* and allows for making unbiased comparisons in the factor scores across groups. When conducting similar analyses on continuous data, the scalar invariance step holds the item intercepts constant. Hirschfeld and Von Brachel [60] and Svetina *et al.* [61] provide worked examples for conducting measurement invariance for both continuous and ordinal data. Measurement invariance testing can include an additional fifth step that fixes the residuals for each item across the groups to test for residual invariance. We did not pursue this step because it is not necessary for making comparisons across groups [62]. We adapted Rocabado's [33] code for the lavaan [63] package to run our measurement invariance analyses.

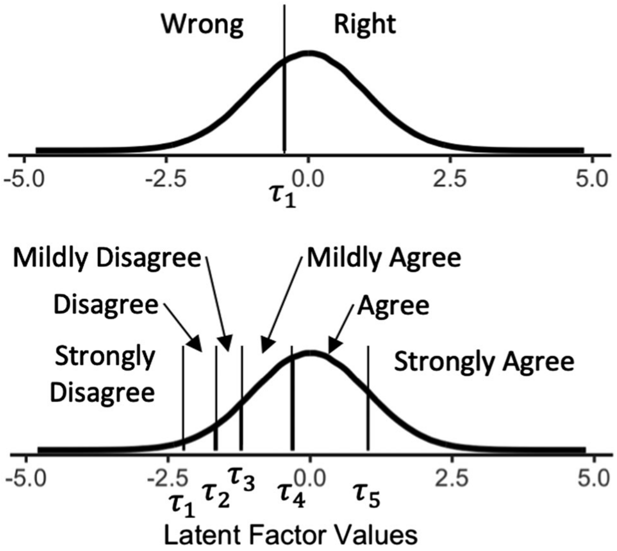

FIG. 2. Latent continuous-response variables for a binary item and a six-point, Likert-scale item. The thresholds (τ) link the observed ordered categorical responses to the latent continuous-response variables. The variable with two categories has one threshold and the variable with six response categories has five thresholds.

Across each step in our analyses, we used three fit indices to determine model fit: the root mean square error of approximation (RMSEA), the comparative fit index (CFI), and the Tucker-Lewis index (TLI). To interpret the CFI and TLI, we used two cutoffs of >0.90 and >0.95, with the >0.95 providing a more conservative cutoff. For

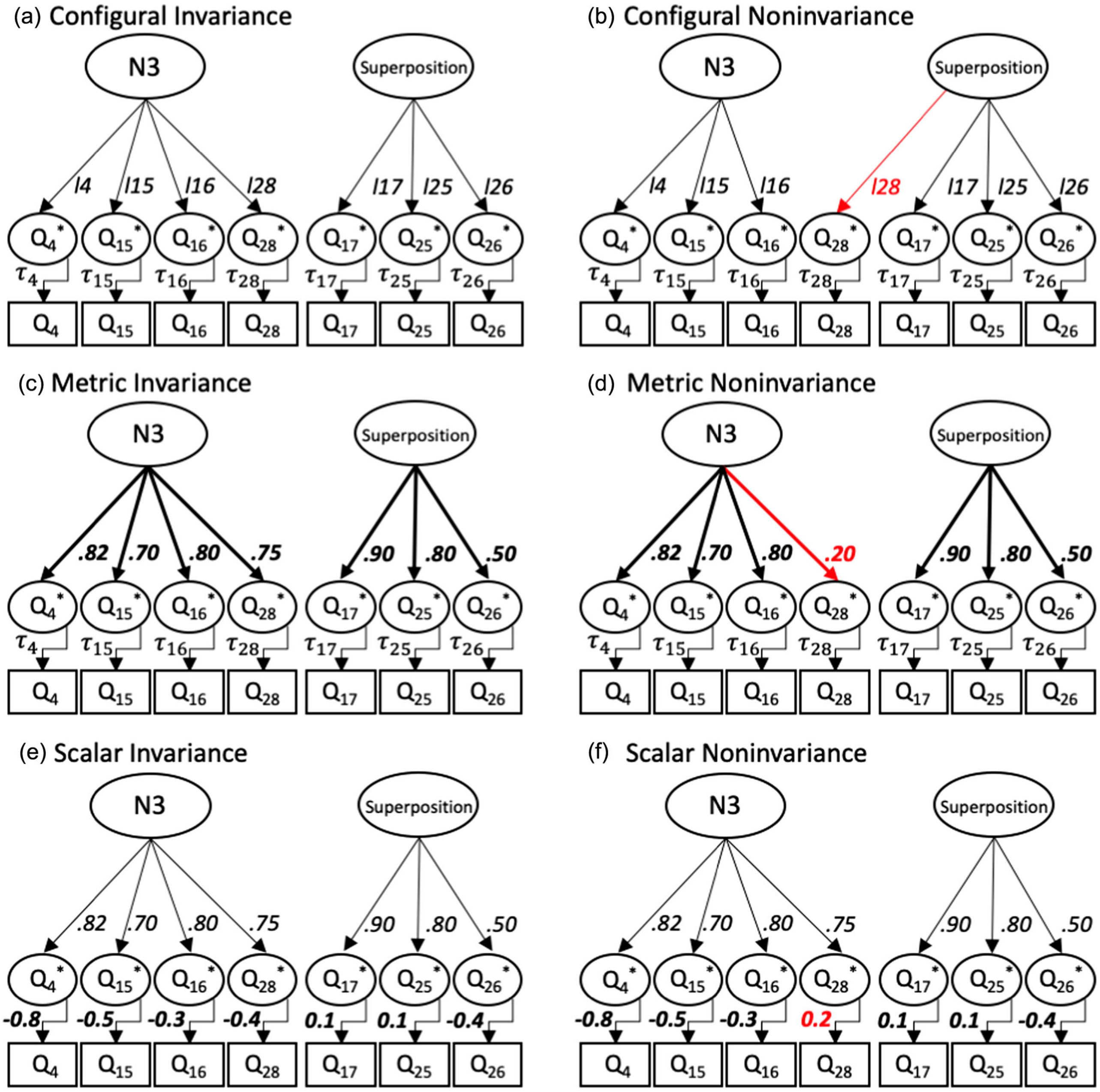
(a) Configural Invariance
N3
Superposition
l4
l15
l16
l28
l17
l25
l26
Q_4^* Q_{15}^* Q_{16}^* Q_{28}^* Q_{17}^* Q_{25}^* Q_{26}^*
τ_4 τ_{15} τ_{16} τ_{28} τ_{17} τ_{25} τ_{26}
Q_4 Q_{15} Q_{16} Q_{28} Q_{17} Q_{25} Q_{26}
(b) Configural Noninvariance
N3
Superposition
l4
l15
l16
l28
l17
l25
l26
Q_4^* Q_{15}^* Q_{16}^* Q_{28}^* Q_{17}^* Q_{25}^* Q_{26}^*
τ_4 τ_{15} τ_{16} τ_{28} τ_{17} τ_{25} τ_{26}
Q_4 Q_{15} Q_{16} Q_{28} Q_{17} Q_{25} Q_{26}
(c) Metric Invariance
N3
Superposition
.82
.70
.80
.75
.90
.80
.50
Q_4^* Q_{15}^* Q_{16}^* Q_{28}^* Q_{17}^* Q_{25}^* Q_{26}^*
τ_4 τ_{15} τ_{16} τ_{28} τ_{17} τ_{25} τ_{26}
Q_4 Q_{15} Q_{16} Q_{28} Q_{17} Q_{25} Q_{26}
(d) Metric Noninvariance
N3
Superposition
.82
.70
.80
.20
.90
.80
.50
Q_4^* Q_{15}^* Q_{16}^* Q_{28}^* Q_{17}^* Q_{25}^* Q_{26}^*
τ_4 τ_{15} τ_{16} τ_{28} τ_{17} τ_{25} τ_{26}
Q_4 Q_{15} Q_{16} Q_{28} Q_{17} Q_{25} Q_{26}
(e) Scalar Invariance
N3
Superposition
.82
.70
.80
.75
.90
.80
.50
Q_4^* Q_{15}^* Q_{16}^* Q_{28}^* Q_{17}^* Q_{25}^* Q_{26}^*
-0.8
-0.5
-0.3
-0.4
0.1
0.1
-0.4
Q_4 Q_{15} Q_{16} Q_{28} Q_{17} Q_{25} Q_{26}
(f) Scalar Noninvariance
N3
Superposition
.82
.70
.80
.75
.90
.80
.50
Q_4^* Q_{15}^* Q_{16}^* Q_{28}^* Q_{17}^* Q_{25}^* Q_{26}^*
-0.8
-0.5
-0.3
0.2
0.1
0.1
-0.4
Q_4 Q_{15} Q_{16} Q_{28} Q_{17} Q_{25} Q_{26}

FIG. 3. An explanatory model of measurement invariance testing using path diagrams for two factors on the FCI, see Table II, using simulated data. Measurement invariance testing checks that this structure is consistent across groups by constraining different values across each step. Configural invariance testing constrains the paths of the factor loadings, metric invariance testing constrains the factor loading values, and scalar invariance testing constrains the item thresholds. The bold outputs are the constraints held constant for all groups for the respective level of invariance testing. The invariant models (A, C, and E) show models with constrained values. The noninvariance models (B, D, and F) depict failure at each level of measurement invariance testing. The point of failure is denoted in red and is always for item Q28. For configural noninvariance, an item loaded on a different factor than the constrained model. For metric noninvariance, a factor loading differed from the constrained model. For scalar noninvariance, the item threshold differed. Figure adapted from Putnick and Bornstein [38], Flora [30], and Bowen and Masa [59].

RMSEA, we used a cutoff of <0.05 [38]. To establish metric and scalar invariance, the changes in the fit indices across the steps in our process needed to be $\Delta < 0.02$ for CFI and TLI and $\Delta < 0.015$ for RMSEA [38]. In our first step, we also evaluated the factor loadings for each item, which indicate the amount of variance in that item explained by the factor. We used a minimum factor loading of 0.6 to inform which items or factors may have contributed to differences in the fit indices across the groups [64]. Researchers propose factor loadings of 0.5, explaining 25%

of the variance in the item, as an absolute minimum and 0.7 as a preferred minimum, with 50% explained variance [64].

Researchers have three options if the model does not pass the fit index cutoff criteria or delta cutoff criteria. First, researchers can perform a partial invariance test by either sequentially releasing or adding constraints for individual items and retesting the model until the fit indices and delta cutoff criteria are satisfied. Second, researchers can drop items with noninvariant loadings and restart the measurement invariance analysis. Finally, researchers can conclude that the model is noninvariant across the groups.

In the Appendix, we discuss fit statistics, estimators, cutoff criteria for measurement invariance testing, and the role of the type of data in the methods we used in this study.

E. Reliability

Researchers frequently measure reliability using Cronbach's alpha or similar statistics [30] such as omega which does not have the assumption of tau equivalence that Cronbach's alpha has. Tau equivalence is indicated in a CFA model by equal factor loadings for all items in a factor. These measures represent the signal-to-noise ratio as the ratio of the total score variance explained by the factor divided by the estimated variance of the observed total score. Flora [30] provides a detailed tutorial on reliability analysis for continuous, categorical, and hierarchical factors.

We used the semTools [65] and lavaan [63] packages to calculate omega. We could choose between calculating omega based on the underlying CFA model or based on the observed ordinal scale data. Because we suspected that most subsequent work applying the EW5M model will use the observed ordinal scale by taking the sum of correct answers, such as is common for the FCI overall score, we calculated reliability for the factors using the observed ordinal scale. The semTools package documentation provides guidance for calculating ω and details on how it was calculated.

F. Factor models

Because two versions of the Eaton and Willoughby five-factor model exist [43], we had to decide which model to use. We conducted a CFA for the general best fit on the EW5 and the EW5M models for the pretest and post-test across all ten social identity groups. Because the EW5 model fits the data better, with higher TLI and CFI for 19 of the 20 analyses than the EW5M model, we focus on the findings from the EW5 model.

G. Descriptive statistics

We included descriptive statistics for the mean, standard deviation, and sample size (N) with the fit indices in our initial models (step 1). These descriptive statistics provided insight into the relationship between the groups' performance on the FCI and the fit indices for the EW5 model. We took this step in response to Planinic *et al.* [54] finding that the FCI functioned differently for groups of students with lower scores and the items on factors 1 and 2 tending to be easier and have higher rates of guessing [53,54].

VII. FINDINGS

We begin by discussing the initial CFA results, step 1 in the measurement invariance testing, for the pretest and post-test across the ten social identity groups. The findings then present the results for configural, metric, and scalar invariance results, steps 2–4.

A. Step 1: Initial models

The CFA for each social identity group found that all groups met the <0.05 cutoff for RMSEA and the less

TABLE IV. Model fit indices and descriptive statistics across race and gender for the pretests and post-tests. We used cutoffs for acceptable fits of >0.95 for CFI and TLI and <0.05 for RMSEA. All fit indices are robust metrics.

		White		Black		Asian		Hispanic		White Hispanic	
Time	Statistic	M	W	M	W	M	W	M	W	M	W
Pre	CFI$_R$	0.97	0.93	0.96	0.94	0.98	0.97	0.93	0.97	0.96	0.91
Pre	TLI$_R$	0.97	0.92	0.96	0.94	0.97	0.97	0.92	0.96	0.95	0.90
Pre	RMSEA$_R$	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.02	0.03	0.03
Pre	Mean (%)	45.8	29.9	31.7	25.4	44.5	36.6	31.8	25.0	37.4	26.9
Pre	S.D. (%)	21.9	17.6	19.4	17.7	24.3	23.7	16.7	16.3	19.3	16.6
Pre	N	5228	3803	261	351	754	754	455	309	537	378
Post	CFI$_R$	0.98	0.98	0.97	0.97	0.99	0.98	0.97	0.98	0.98	0.98
Post	TLI$_R$	0.98	0.97	0.97	0.96	0.99	0.97	0.96	0.98	0.98	0.98
Post	RMSEA$_R$	0.03	0.03	0.03	0.03	0.03	0.04	0.03	0.03	0.03	0.03
Post	Mean (%)	66.4	52.8	45.3	41.6	61.8	56.6	49.4	42.8	58.0	47.1
Post	S.D. (%)	23.4	25.0	23.2	23.9	26.8	27.2	22.4	25.2	24.9	24.5
Post	N	4479	3420	191	270	612	628	360	248	480	325

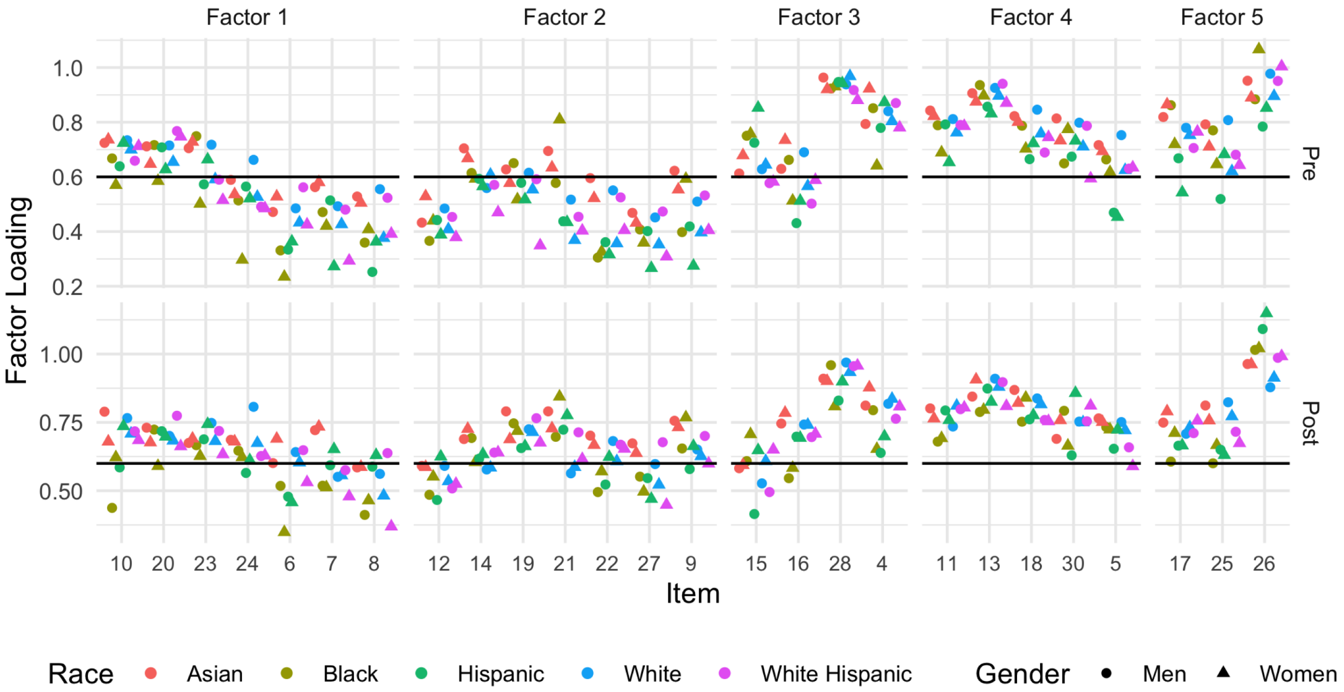

FIG. 4. The factor loading for each item across the ten social identity groups organized by factor for the pretest and post-test. The black horizontal line represents the 0.6 value we applied as a cutoff.

conservative >0.90 cutoff for CFI and TLI, as shown in Table IV. Of the 20 groups, 16 met the more conservative cutoff of >0.95 for CFI and TLI. The four CFAs below the conservative cutoff were on the pretest for Black women, Hispanic men, White women, and White Hispanic women. All groups met the more conservative >0.95 cutoff for CFI and TLI on the post-test. As we discuss later, the higher CFI and TLI on the post-test for every group likely resulted from the learning that occurred during the course.

The standardized factor loading for each item, shown in Fig. 4 and in the Appendix in Table VI, also provided evidence that the EW5 model fit the data well. Some of the factor loadings, however, were below 0.6. Far fewer items had any factor loadings below 0.6 on the post-test (15 items and 57 loadings across the 10 CFAs) than on the pretest (20 items and 118 loadings across the 10 CFAs). As shown in Fig. 4, most items with factor loadings less than 0.6 occurred on factors 1 and 2 for Newton's first and second law, which had 14 items between the two factors, for both the pretest (14 items) and the post-test (12 items).

Factor loadings less than 0.6 were distributed across race and gender groups on the pretest and post-test. On the pretest, the number of items with low factor loadings ranged from 7 for Asian men to 15 for Hispanic men and White Hispanic women. On the post-test, the number of items with low factor loadings ranged from 3 for Asian men to 10 for Hispanic men. On the post-test, items 6, 7, 8, 12, 15, and 27 had factor loadings less than 0.6 for five or more groups with five of these six items loading on factor 1 or 2.

Because several of the pretests fell below our conservative threshold and findings by Planinic *et al.* [54] that the FCI may function differently for students with low performance, we explored the relationships

TABLE V. Measurement invariance test results for pretest and post-test scores. We used cutoffs for acceptable Δ values of <0.02 for CFI and TLI and <0.015 for RMSEA. All fit indices are robust metrics.

Steps	CFI_R	TLI_R	RMSEA_R	Δ calculation	$\Delta\mathrm{CFI}_R$	$\Delta\mathrm{TLI}_R$	$\Delta\mathrm{RMSEA}_R$
Pretest							
S2 configural	0.954	0.949	0.031	...	...	...	...
S3 metric	0.957	0.956	0.029	S2–S3	−0.004	−0.007	0.002
S4 scalar	0.958	0.955	0.031	S3–S4	−0.001	0.001	−0.002
Post-test							
S2 configural	0.979	0.977	0.029	...	...	...	...
S3 metric	0.980	0.979	0.027	S2–S3	−0.001	−0.003	0.002
S4 scalar	0.977	0.976	0.032	S3–S4	0.003	0.003	−0.005

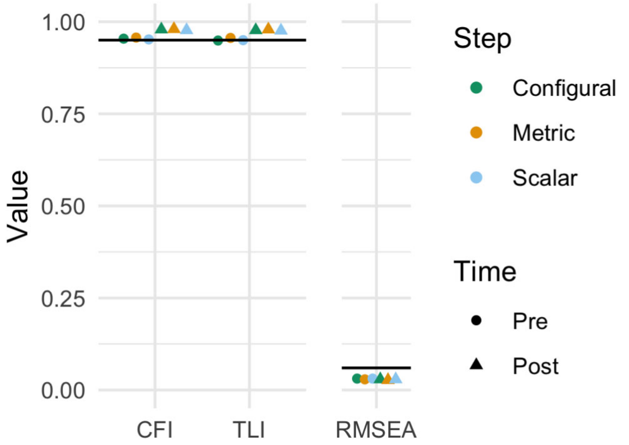

FIG. 5. The index values for configural, metric, and scalar invariance. CFI and TLI were at or above the 0.95 cutoff (shown as a black line) except for the pretest TLI for configural invariance (0.949) and all RMSEA values were below the 0.05 cutoff.

between performance and model fit. While a full investigation of these limitations is beyond the scope of this paper, Table IV provides the mean, standard deviation, and N for each group. The four analyses with CFI below the 0.95 cutoff had mean FCI scores ranging from 25% to 32%. The 16 analyses with CFI greater than 0.95 had scores ranging from 25% to 66%. These results tentatively indicate that the EW5 model may consistently fit the FCI data well for datasets with a mean score above about 35%.

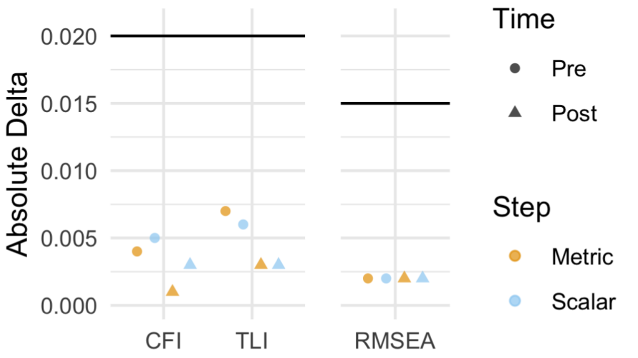

FIG. 6. The absolute value of the delta from configural to metric and from metric to scalar for each of the fit indices. CFI and TLI were all below the 0.02 cutoff and RMSEA were all below the 0.015 cutoff.

Because these initial models, particularly the post-test models, showed adequate fit, we moved on to the subsequent steps of testing measurement invariance.

B. Steps 2–4: Configural, metric, and scalar invariance

Table V and Figs. 5 and 6 show the fit indices for steps 2–4 of the measurement invariance analysis for both the pretest and post-test. The model passed each of these steps for both the pretest and post-test using the conservative cutoffs for CFI (> 0.95), RMSEA (<0.05), and TLI except for the 0.949 TLI on the pretest, see Fig. 5. Deltas from step 2 to 3 and step 3 to 4 were within cutoffs of <0.02 for CFI

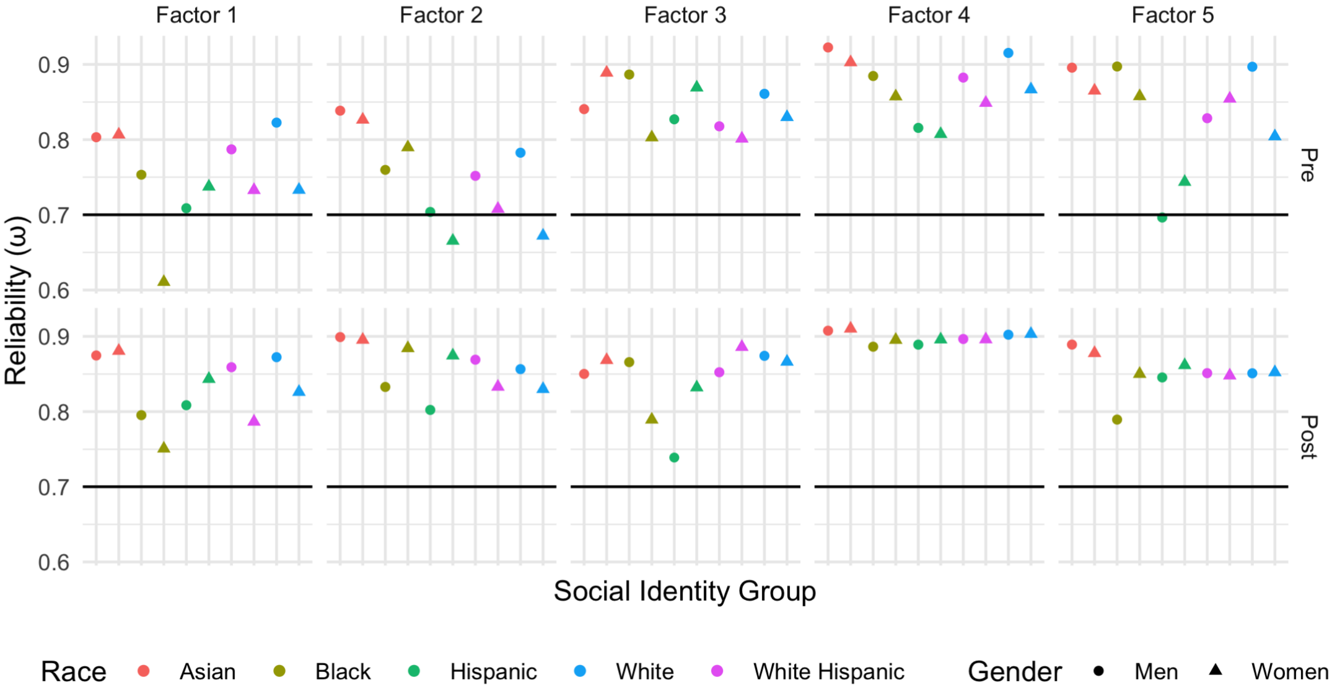

FIG. 7. The reliability (ω) for each of the factors and the total scores on the pretest and post-test across all ten social identity groups. The black horizontal line at 0.7 marks the commonly accepted minimum value for reliability. Reliabilities exceeded the 0.7 cutoff for 46 of the 50 pretest measures and for all of the post-test measures.

and TLI and <0.015 for RMSEA on both the pretest and post-test, see Fig. 6. These results indicate measurement invariance of the EW5 model for the FCI. Configural invariance indicates that the same factor structure fits the data well for all ten social identity groups [38]. Metric invariance indicates that each item contributes to each factor in similar ways across all groups [38]. Scalar invariance indicates that mean differences in the latent factor capture all mean differences in the shared variance of the items on the FCI similarly across all ten social identity groups [38].

C. Reliability

Figure 7 shows the reliability of each factor on the pretest and post-test for each of the ten social identity groups. Of the 100 reliability scores for the factors, 4 fell below the 0.7 cutoff indicated by the black line in Fig. 7; all 4 were on the pretest. Factors 1 and 2 tended to have lower reliability on the pretest than all of the other reliability statistics. This lower reliability for factors 1 and 2 on the pretest was consistent with the frequent poor factor loadings for items included in these factors presented earlier in the results.

VIII. DISCUSSION

The measurement invariance and reliability indicate the EW5 factor model for the FCI performs similarly across the ten social identity groups. Intersectionality theory posits that differences can exist across intersectional groups and that research needs to account for this possibility lest it hides and perpetuates inequities. Extensive prior research [6,9,19,66–69], including research across the intersections of gender and race [15], has identified large differences in both pretest and post-test scores on the FCI and other research-based assessments. Finding measurement invariance and reliability indicates those results for the FCI represent real differences in content knowledge and not artifacts of how the test or items functioned for different social identity groups. This contrasts the findings of Traxler *et al.* [8] that some items on the FCI functioned differently for men and women, which we discuss further below.

The post-test measures of invariance for all social identity groups met the fit indices' thresholds and indicate the FCI measures the EW5 factors similarly across all groups in the analysis as shown in Tables IV and V. The pretest showed weaker performance than the post-test, but all fit indices were acceptable. Meeting configural, metric, and scalar invariance indicates that across all of the groups, the same factor structure fits the data well. It also indicates that each item contributes to each factor similarly across groups and the latent factors were measured on the same scale across groups. These results indicate that EW5 model of the FCI collects comparable data across these ten social identity groups. Instructors and researchers need not limit their investigations of equity to overall scores as they can compare group performance across each of the EW5 factors. This more specific analysis can indicate if differences across groups disproportionately occur for a specific topic that instruction can then address.

The factor loadings across the ten social identity groups for the pretest and post-test also indicate that limitations of the internal structure validity of the EW5 model and FCI are issues across all ten social identity groups and not unique to any one group. The low factor loadings, < 0.6, tended to occur on factors 1 and 2 on the pretest and to a lesser extent on the post-test.

The factors on the EW5 model had acceptable reliability in 96 of the 100 evaluations. Consistent with the lower factor loadings for factors 1 and 2 on the pretest, the reliability for these factors on the pretest tended to be worse than for the other factors on the pretest and all five factors on the post-test. The poorer reliability on the pretest follows from fewer students knowing the right answer and a larger proportion of correct answers resulting from guessing. Since the factors are measuring conceptual knowledge, the signal will be lower on the pretest than on the post-test as the scores tended to increase by approximately one standard deviation. The factors cannot explain the variance created from correct answers due to guessing as it will be very weakly correlated across the items. Thus, the reliability will tend to be lower on pretests, particularly for items with higher guessing rates.

A. Applying the EW5 model to prior research

While measurement invariance indicates instructors and researchers can compare scores on the five factors of the EW5 model, the results also indicate that factors 1 and 2 perform poorly, especially for groups with lower scores. Lower scores occur more on a pretest than on a post-test. Lower scores also occur for students exposed to systemic racism, sexism, or class oppression, such as through the underfunding of schools in poor or majority non-White neighborhoods leading to fewer physics courses or less prepared physics instructors. Thus, while the EW5 model can parse additional data from the FCI, the low factor loadings, many of the items for Newton's first and second laws and kinematics (factors 1 and 2) limit the ability of the FCI to inform interventions or research directed at understanding or addressing inequities for students entering college physics courses. While identifying the specific problematic items or the causes of their limitations lies beyond the analytical scope of this article, we can apply the EW5 model to prior studies of the FCI to further understand these limitations.

As we discussed in the literature review, Sec. V, Wang and Bao [53] used a three-parameter item response theory model to investigate the performance of the items on the FCI. Their results indicated that the items loading on factors 1 and 2 for Newton's first and second laws tended to be easier than the other items and tended to have higher rates of guessing, where guessing is the likelihood that an individual with very little Newtonian knowledge would get an item correct. Of the 14 items on factors 1 and 2, 7 items had guessing rates greater than 20%. Four of these items, 6, 7, 8, and 27, also had poor factor loadings on the pretest in our results. Wang and colleagues' findings are consistent with our findings that when large proportions of students have low scores on these factors noise due to correct answers from guessing may obscure the signal from correct answers representing understanding the concepts.

Traxler *et al.* [8] found that ten items on the FCI functioned differently for men and women, which indicates that these items may be biased in favor of men. Nine of these ten items are also items that load on factors 1 and 2 for Newton's first and second law in the EW5 model. These results align with our findings that the items on factors 1 and 2 provide poorer quality information (i.e., low signal to noise ratios). Their results also contrast with our own findings. While our results indicated the EW5 model of the FCI functioned similarly across the intersections of race or ethnicity and gender, Traxler *et al.* [8] found these items were biased against women and exaggerated the gender differences in conceptual understanding of Newtonian physics. Additional work to determine if these items are biased against women or if they function differently because of different levels of knowledge across men and women and to extend that work across the intersection of race and gender can further inform the limitations and usefulness of the FCI.

IX. CONCLUSIONS

By analyzing the FCI using the EW5 model, researchers and instructors can extract additional information from the FCI about inequities in a course and a course or pedagogy's impacts on equity across the intersections of race and gender. The items covering Newton's first and second law, however, provide more limited information in settings where most students have little formal knowledge of Newtonian mechanics. Our results indicated that the EW5 model consistently fit the FCI data well for datasets with a mean score above about 35%, and we found acceptable fits for scores as low as 25%. Studies seeking to apply the EW5 model in contexts with scores below approximately 35% will need to take additional steps to ensure the factors are measuring student understanding rather than correct answers due to guessing.

The FCI has supported the physics education community in understanding physics learning and developing more effective pedagogies. This study, however, shows the limitations of the FCI in providing a finer-grained measure of students' Newtonian knowledge as a pretest or in contexts where students tend to have lower scores. A next generation of research-based assessment could provide instructors with these finer-grained measures across a wider range of knowledge levels to better guide instruction and better inform the development of research-based instructional strategies.

X. LIMITATIONS

Our study had several limitations that we tried to mitigate. While the institutions in our dataset were more diverse than most physics education research publications [70], the dataset was not fully representative of institutions in the United States [48,71]. Our results indicate that measurement invariance was met across our multi-institutional sample, but it is possible that we would find different results had we analyzed a more diverse sample. The poorer performance of the FCI for students with low Newtonian knowledge indicates it may not work well as a pretest in a high school or college course where few students have had prior physics instruction.

For confirmatory factor analysis using the WLSMV estimator, research recommends a sample size of at least 200 persons. For this reason, we could not study scores for individuals who were multiracial, had nonbinary genders, or identified as part of a less-well-represented racial group. Our study had enough power to analyze scores for five racial groups across two genders. While we found evidence of measurement invariance for the EW5 model on the FCI, these findings may not generalize for all groups. Nor do these results generalize to the translated versions of the FCI.

While this study was grounded in a desire to support intersectional research, it was not itself fully intersectional [37]. The nature of measurement invariance testing only allows for the inclusion of discrete groups with large enough sample sizes. Measurement invariance testing excludes individuals from smaller subgroups in the model (e.g., Black, Hispanic, and transgender students). Our analysis was limited in regard to power structures and only included a limited investigation of students in algebra- and calculus-based physics courses due to sample size constraints.

ACKNOWLEDGMENTS

This work was made possible by funding from the National Science Foundation (Grants No. 1928596 and No. 2141847). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. This work would not have been possible without the effort and intellectual contribution of Alicen Morley.

APPENDIX A: ABSOLUTE FIT INDICES

The chi-squared statistic (χ^2) measures the difference between the observed data and what we would expect if there was no relationship between the data and the FCI's factors. A significant χ^2 will show a difference between the observed and expected samples approaching zero [72,73].

While χ^2 is the traditional measure of model fit and still holds popularity as a fit statistic [72–75], it has significant limitations. It relies on the data fitting a normal curve and is sensitive to severe deviations from the curve [76]. χ^2 is also very sensitive to sample size. When datasets are large, χ^2 almost always rejects the model, and when the sample is small, it cannot distinguish between good-fitting models and poor-fitting models [77]. We did not rely on chi-squared results to determine model fit because our data are binary and several of the groups we analyzed were near the 200-person minimum.

The standardized root mean square residual (SRMR) is another fit statistic that describes the absolute fit of a model. SRMR examines the average difference between the observed and implied covariance matrices to contrast the predicted and observed samples. This index reports a range from 0 to 1. A good fit is less than 0.08, while a well-fitting model is considered to be 0.05 or less [75,78,79].

While SRMR is considered a relatively stable fit measure, it will be artificially lower when there is a high number of parameters in the model and for models with large sample sizes [74]. It is also not considered to hold much explanatory power because of the nature of averages. For example, the same SRMR score can be achieved with a few large differences from the implied model or with many small differences. For this reason, we do not examine the SRMR or report it.

The root mean squared error of approximation (RMSEA) is the most widely used absolute fit index because it is sensitive to the number of degrees of freedom (d.o.f.) in the model, essentially rewarding models with larger d.o.f. It favors parsimony by choosing a model with fewer parameters to fit the sample's covariance matrix [78]. For example, it can be calculated by

$$\mathrm{RMSEA} = \sqrt{\frac{\tilde{F}_T}{df_T}},$$

where the minimized fit function ($\tilde{F}$) and the degrees of freedom are for the model being tested [80].

RMSEA cutoffs changed significantly since the early 1990s when a value in the 0.05–0.10 range was considered fair, with values above 0.10 being poor [81]. More recently, a cutoff of 0.06 or a strict upper limit of 0.07 is the consensus [75]. However, instrument bias research tends to cut off RMSEA at 0.05 [31,38]. For this reason, we employ a cutoff of 0.05 for this work.

APPENDIX B: RELATIVE GOODNESS OF FIT INDICES

Relative goodness of fit indices compare the minimized fit function for the tested model with a baseline model, which is the model with the worst fit where all covariances are set to zero and the variances are freely estimated [31]. We employ two relative fit indices for this work, the comparative fit index (CFI) and the Tucker-Lewis index (TLI).

The CFI performs well even when the sample size is small [82]. First introduced by Bentler [83], the CFI assumes a baseline model where all latent variables are uncorrelated and compares the sample's covariance matrix for the tested model with the baseline model. Values range from 0.0 to 1.0, with values closer to 1.0 indicating a good fit. CFI ≥ 0.90 has been a common cutoff criterion [74]. Hu and Bentler [75] recommended a minimum CFI of 0.95 to ensure that misspecified models are not accepted. An example of CFI calculation is

$$\mathrm{CFI} = 1 - \frac{\tilde{F}_T}{\tilde{F}_B},$$

where the ratios of the minimized fit functions of the tested and baseline models determine the CFI.

The Tucker-Lewis Index (TLI) is a good indicator for smaller datasets [84]. The TLI compares the baseline and tested models while accounting for the degrees of freedom in each model. A larger TLI value indicates a better fit with a common cutoff criterion of 0.95 [75]. TLI can be calculated as

$$\mathrm{TLI} = 1 - \frac{\tilde{F}_T}{\tilde{F}_B} * \frac{df_B}{df_T}.$$

APPENDIX C: ESTIMATORS

Fit indices are based on a fit function specific to a given estimation method. Which estimation method to use depends on the structure of the data and the proposed latent factors. Research on measurement invariance often provides guidance for continuous data with large sample sizes, comparing only two groups, or models with few factors [33,38]. Using the default settings in SEM packages such as lavaan can produce misleading results due to the sensitivity of fit indices to estimator choice.

The lavaan package uses maximum likelihood (ML) as the default estimator. ML assumes continuous data and is sensitive to small sample sizes [80]. Applying ML to the covariance matrix of noncontinuous data can lead to biased parameter estimates, inaccurate standard errors, and a misleading chi-squared statistic, [80] [p. 410]. While the ML Robust (MLR) estimator can work well for binary data, like correct or incorrect item scores, it is sensitive to sample size. Also, lavaan is limited in its categorical ML estimation capabilities [85].

Unweighted least squares (ULS) and diagonally weighted least squares (DWLS) are appropriate for categorical data and are available in the lavaan package. These estimators, however, are sensitive to sample size, model nesting, and the number of factors in the model [80]. Xia and Yang [80] also found that using common cutoffs for RMSEA, CFI, and TLI with ULS and DWLS estimators can lead to accepting models with misfit. Xia and Yang [80] advise that researchers do not use these fit indices as go or no-go tests but instead use them to inform an iterative approach to model evaluation.

To accommodate the EW5 models, small sample sizes, and binary data, we employed the weighted least squares mean and variance adjusted (WLSMV) estimator. WLSMV is the "robust" version of the DWLS estimator [63], meaning that it will account for outliers and reduce the likelihood of accepting a misspecified model. Its recommended threshold of 200 cases per group or sample is lower than most other estimators [58,86], though it tends to over-reject models at that sample size [87].

APPENDIX D: CHANGE IN FIT INDICES ACROSS STEPS IN MEASUREMENT INVARIANCE TESTING

For measurement invariance testing, researchers often examine the size of the change in fit indices from one step to another. Putnick and Bornstein [38] recommend any change in the CFI larger than 0.02 to fail measurement invariance testing. Eaton [51] uses a cutoff of 0.01 for the CFI and 0.015 for RMSEA; Cheung and Rensvold [88] recommend these cutoffs for large, relatively even sample sizes. Rutkowski and Svetina [89] support changes of 0.02 for the CFI and 0.03 for RMSEA at the metric invariance level and 0.01 for CFI and 0.01 for RMSEA is appropriate for the scalar invariance step. Chen [90] recommends that ΔCFI should be no more than 0.005 and ΔRMSEA should be 0.010 or smaller based on unequal sample sizes and each sample is less than 300.

Research has not yet reached a consensus about the best-fit indices or cutoff values under all conditions. Finding recommendations that met the unique circumstances of our data was difficult due to the large difference in sample size between our largest (White Men; $n = 6200$) and smallest group (Hispanic women; $n = 232$). All of the studies discussed above-examined groups of comparable size except for Chen [90], which did not use binary data or WLSMV. No single recommendation satisfied all of our study's characteristics. For this reason, we used the fit indices and cutoffs as a guide and not as go or no-go tests.

We looked at both cutoffs of 0.90 and 0.95 for CFI and we evaluated RMSEA for a cutoff of 0.05. We also examined the change in CFI, TLI, and RMSEA from step 1 (configural) to step 2 (metric) and step 2 to step 3 (scalar). Our analysis also treats the delta CFI and RMSEA scores from one step of measurement invariance testing to another with caution. We explored alternative models by evaluating the factor loadings and using the modIndices function in lavaan to identify the extent to which changing the factor structure would improve the model fits.

APPENDIX E: DATA PREPARATION

We removed data for students who took less than 5 min or did not complete all of the questions. Completing all of the questions was required for the measurement invariance testing. Of the students who answered any questions on the pretest or post-test and consented to share their data, 81.1% answered questions on the pretest and 68.7% answered questions on the post-test. For the students who answered any questions on the pretest, 94.4% of the data was retained with 3.9% removed due to time, 1.4% removed due to completeness, and 0.3% removed due to both. For the students who answered any questions on the post-test, 89.9% of the data was retained with 7.4% removed due to time, 2.3% removed due to completeness, and 0.4% removed due to both.

APPENDIX F: FACTOR LOADINGS

Table VI provides the factor loadings for the initial CFAs on the pretest and post-test for all ten social identity groups.

TABLE VI. Factor loadings across the ten social identity groups for the pretest and post-test. Note that in contrast to the natural sciences, which usually deal with observable constructs that can be measured using a single instrument (e.g., measuring length with a ruler), variables in psychology are nonobservable (latent), and can be measured using several facets or indicators. Alpha coefficient is lower bound to omega coefficient. They are equal if and only if the items fit the single-factor model with equal factor loading.

		Pretest CFI values										Post-test CFI values									
		White		Black		Asian		Hispanic		White Hispanic		White		Black		Asian		Hispanic		White Hispanic	
Factor	question	M	W	M	W	M	W	M	W	M	W	M	W	M	W	M	W	M	W	M	W
1	6	0.48	0.43	0.33	0.23	0.47	0.53	0.33	0.36	0.56	0.43	0.64	0.60	0.52	0.35	0.60	0.69	0.48	0.46	0.65	0.53
1	7	0.49	0.43	0.47	0.42	0.56	0.58	0.51	0.27	0.48	0.29	0.55	0.56	0.52	0.51	0.72	0.73	0.59	0.65	0.58	0.48
1	8	0.56	0.38	0.36	0.41	0.53	0.51	0.25	0.36	0.52	0.39	0.56	0.48	0.41	0.46	0.59	0.59	0.59	0.63	0.64	0.37
1	10	0.73	0.70	0.67	0.57	0.72	0.74	0.64	0.72	0.66	0.71	0.77	0.71	0.44	0.62	0.79	0.68	0.59	0.74	0.72	0.68
1	20	0.72	0.65	0.72	0.58	0.71	0.65	0.71	0.63	0.77	0.75	0.70	0.68	0.72	0.59	0.73	0.68	0.72	0.70	0.77	0.66

(Table continued)

TABLE VI. *(Continued)*

		Pretest CFI values										Post-test CFI values									
		White		Black		Asian		Hispanic		White Hispanic		White		Black		Asian		Hispanic		White Hispanic	
Factor	question	M	W	M	W	M	W	M	W	M	W	M	W	M	W	M	W	M	W	M	W
1	23	0.72	0.59	0.75	0.50	0.71	0.73	0.57	0.66	0.59	0.52	0.75	0.68	0.67	0.63	0.67	0.69	0.69	0.74	0.72	0.63
1	24	0.66	0.53	0.51	0.30	0.59	0.54	0.56	0.52	0.49	0.49	0.81	0.67	0.65	0.62	0.69	0.68	0.57	0.61	0.63	0.63
2	9	0.51	0.40	0.40	0.59	0.62	0.55	0.42	0.27	0.53	0.41	0.65	0.63	0.65	0.77	0.76	0.73	0.58	0.66	0.70	0.60
2	12	0.48	0.41	0.37	0.44	0.43	0.53	0.44	0.39	0.45	0.38	0.59	0.53	0.49	0.55	0.59	0.59	0.47	0.62	0.51	0.52
2	14	0.56	0.61	0.61	0.59	0.70	0.67	0.59	0.57	0.57	0.47	0.58	0.58	0.69	0.60	0.69	0.73	0.62	0.63	0.64	0.64
2	19	0.62	0.55	0.65	0.52	0.63	0.58	0.58	0.52	0.59	0.35	0.72	0.71	0.75	0.72	0.79	0.69	0.66	0.66	0.77	0.68
2	21	0.52	0.37	0.58	0.81	0.69	0.63	0.44	0.43	0.45	0.40	0.56	0.59	0.70	0.84	0.79	0.73	0.72	0.78	0.71	0.62
2	22	0.55	0.36	0.30	0.33	0.60	0.52	0.36	0.32	0.53	0.41	0.68	0.61	0.50	0.57	0.70	0.67	0.52	0.62	0.67	0.65
2	27	0.45	0.35	0.41	0.36	0.47	0.43	0.40	0.27	0.47	0.31	0.60	0.52	0.55	0.50	0.67	0.64	0.55	0.47	0.68	0.45
3	4	0.84	0.80	0.85	0.64	0.79	0.92	0.78	0.87	0.87	0.78	0.82	0.84	0.79	0.65	0.81	0.88	0.64	0.70	0.76	0.81
3	15	0.63	0.64	0.75	0.76	0.61	0.68	0.72	0.85	0.58	0.58	0.53	0.61	0.61	0.71	0.58	0.59	0.41	0.65	0.50	0.65
3	16	0.69	0.57	0.66	0.51	0.63	0.73	0.43	0.51	0.50	0.59	0.74	0.74	0.55	0.58	0.75	0.78	0.70	0.69	0.70	0.71
3	28	0.94	0.97	0.92	0.93	0.96	0.92	0.95	0.94	0.92	0.88	0.97	0.93	0.96	0.81	0.91	0.90	0.83	0.90	0.96	0.96
4	5	0.75	0.62	0.66	0.62	0.72	0.69	0.47	0.45	0.63	0.63	0.75	0.72	0.73	0.72	0.76	0.75	0.65	0.72	0.66	0.59
4	11	0.81	0.76	0.79	0.69	0.84	0.82	0.79	0.65	0.79	0.79	0.73	0.81	0.68	0.69	0.80	0.76	0.79	0.76	0.80	0.80
4	13	0.93	0.90	0.94	0.90	0.91	0.88	0.86	0.83	0.94	0.87	0.91	0.88	0.79	0.80	0.84	0.91	0.87	0.83	0.90	0.81
4	18	0.85	0.76	0.79	0.70	0.82	0.80	0.66	0.72	0.69	0.75	0.84	0.82	0.75	0.84	0.87	0.82	0.76	0.78	0.76	0.75
4	30	0.80	0.71	0.65	0.77	0.81	0.73	0.67	0.73	0.79	0.59	0.75	0.75	0.79	0.66	0.69	0.76	0.63	0.86	0.75	0.81
5	17	0.78	0.75	0.86	0.72	0.82	0.86	0.67	0.54	0.71	0.77	0.71	0.73	0.61	0.71	0.75	0.79	0.66	0.67	0.71	0.76
5	25	0.81	0.62	0.77	0.65	0.79	0.71	0.52	0.68	0.68	0.64	0.82	0.77	0.60	0.67	0.81	0.76	0.65	0.63	0.72	0.67
5	26	0.98	0.90	0.88	1.07	0.95	0.89	0.78	0.85	0.95	1.00	0.88	0.91	1.02	1.02	0.96	0.96	1.09	1.15	0.99	0.99

[1] S. Singer and K. A. Smith, Discipline-based education research: Understanding and improving learning in undergraduate science and engineering, J. Eng. Educ. **102**, 468 (2013).

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

[3] A. Madsen, S. B. McKagan, and E. C. Sayre, Resource letter RBAI-1: Research-based assessment instruments in physics and astronomy, Am. J. Phys. **85**, 245 (2017).

[4] B. Van Dusen and J. Nissen, Equity in college physics student learning: A critical quantitative intersectionality investigation, J. Res. Sci. Teach. **57**, 33 (2020).

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

[6] I. Rodriguez, E. Brewe, V. Sawtelle, and L. H. Kramer, Impact of equity models and statistical measures on interpretations of educational reform, Phys. Rev. ST Phys. Educ. Res. **8**, 020103 (2012).

[7] P. Martinková, A. Drabinová, Y.-L. Liaw, E. A. Sanders, J. L. McFarland, and R. M. Price, Checking equity: Why differential item functioning analysis should be a routine part of developing conceptual assessments, CBE Life Sci. Educ. **16**, rm2 (2017).

[8] 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).

[9] A. Traxler and E. Brewe, Equity investigation of attitudinal shifts in introductory physics, Phys. Rev. ST Phys. Educ. Res. **11**, 020132 (2015).

[10] Z. Y. Kalender, E. Marshman, C. D. Schunn, T. J. Nokes-Malach, and C. Singh, Gendered patterns in the construction of physics identity from motivational factors, Phys. Rev. Phys. Educ. Res. **15**, 020119 (2019).

[11] Z. Hazari, G. Sonnert, P. M. Sadler, and M.-C. Shanahan, Connecting high school physics experiences, outcome expectations, physics identity, and physics career choice: A gender study, J. Res. Sci. Teach. **47**, 978 (2010).

[12] T. I. Smith and M. C. Wittmann, Comparing three methods for teaching Newton's third law, Phys. Rev. ST Phys. Educ. Res. **3**, 020105 (2007).

[13] T. I. Smith, M. C. Wittmann, and T. Carter, Applying model analysis to a resource-based analysis of the force and motion conceptual evaluation, Phys. Rev. ST Phys. Educ. Res. **10**, 020102 (2014).

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

[15] B. Van Dusen and J. M. Nissen, Associations between learning assistants, passing introductory physics, and equity: A quantcrit investigation, Phys. Rev. Phys. Educ. Res. **16**, 010117 (2020).

[16] B. Van Dusen and J. Nissen, Equity in college physics student learning: A critical quantitative intersectionality investigation, J. Res. Sci. Teach. **57**, 33 (2019).

[17] J. M. Nissen, I. H. M. Horses, and B. 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).

[18] R. Krakehl and A. M. Kelly, Intersectional analysis of advanced placement physics participation and performance by gender and ethnicity, Phys. Rev. Phys. Educ. Res. **17**, 020105 (2021).

[19] D. Shafer, M. S. Mahmood, and T. Stelzer, Impact of broad categorization on statistical results: How underrepresented minority designation can mask the struggles of both Asian American and African American students, Phys. Rev. Phys. Educ. Res. **17**, 010113 (2021).

[20] C. A. Paul and D. J. Webb, Percent grade scale amplifies racial or ethnic inequities in introductory physics, Phys. Rev. Phys. Educ. Res. **18**, 020103 (2022).

[21] J. Yang, S. DeVore, D. Hewagallage, P. Miller, Q. X. Ryan, and J. Stewart, Using machine learning to identify the most at-risk students in physics classes, Phys. Rev. Phys. Educ. Res. **16**, 020130 (2020).

[22] J. Stewart, G. L. Cochran, R. Henderson, C. Zabriskie, S. DeVore, P. Miller, G. Stewart, and L. Michaluk, Mediational effect of prior preparation on performance differences of students underrepresented in physics, Phys. Rev. Phys. Educ. Res. **17**, 010107 (2021).

[23] A. B. Simmons and A. F. Heckler, Grades, grade component weighting, and demographic disparities in introductory physics, Phys. Rev. Phys. Educ. Res. **16**, 020125 (2020).

[24] Kimberlé Crenshaw, Mapping the margins: Identity politics, intersectionality, and violence against women of color, Stanf. Law Rev. **43**, 1241 (1991), https://www.jstor.org/stable/1229039.

[25] P. H. Collins and S. Bilge, *Intersectionality* (John Wiley & Sons, 2020).

[26] D. Shafer, M. S. Mahmood, and T. Stelzer, Impact of broad categorization on statistical results: How underrepresented minority designation can mask the struggles of both Asian American and African American students, Phys. Rev. Phys. Educ. Res. **17**, 010113 (2021).

[27] Gloria Ladson-Billings, From the achievement gap to the education debt: Understanding achievement in US schools, Educ. Res. **35**, 3 (2006).

[28] J. B Buncher, J. M. Nissen, B. Van Dusen, R. M. Talbot III, and H. Huvard, Bias on the force concept inventory across the intersection of gender and race, presented at PER Conf. 2021, virtual conference, 10.1119/perc.2021.pr.Buncher.

[29] M. T. Kane, Current concerns in validity theory, J. Educ. Measure. **38**, 319 (2001).

[30] D. B Flora, Your coefficient alpha is probably wrong, but which coefficient omega is right? A tutorial on using r to obtain better reliability estimates, Adv. Methods Pract. Psychol. Sci. **3**, 484 (2020).

[31] P. Eaton and S. D. Willoughby, Confirmatory factor analysis applied to the force concept inventory, Phys. Rev. Phys. Educ. Res. **14**, 010124 (2018).

[32] L. Bowleg, When Black + lesbian + woman ≠ Black lesbian woman: The methodological challenges of qualitative and quantitative intersectionality research, Sex Roles **59**, 312 (2008).

[33] G. A. Rocabado, N. A. Kilpatrick, S. R. Mooring, and J. E. Lewis, Can we compare attitude scores among diverse populations? An exploration of measurement invariance testing to support valid comparisons between Black female students and their peers in an organic chemistry course, J. Chem. Educ. **96**, 2371 (2019).

[34] A. Kuhlemeier, Measurement invariance of psychological distress, substance use, and adult social support across race/ethnicity and sex among sexual minority youth, J. Sex Res. **60**, 674 (2023).

[35] N. M. Else-Quest and J. S. Hyde, Intersectionality in quantitative psychological research: II. Methods and techniques, Psychol. Women Q. **40**, 319 (2016).

[36] P. H. Collins, Intersectionality's definitional dilemmas, Annu. Rev. Sociol. **41**, 1 (2015).

[37] S. Cho, K. W. Crenshaw, and L. McCall, Toward a field of intersectionality studies: Theory, J. Women Cult. Soc. **38**, 785 (2013).

[38] D. L. Putnick and M. H. Bornstein, Measurement invariance conventions and reporting: The state of the art and future directions for psychological research, Dev. Rev. **41**, 71 (2016).

[39] Y. Dodge, D. Cox, and D. Commenges, *The Oxford Dictionary of Statistical Terms* (Oxford University Press on Demand, New York, 2003).

[40] J. B. Ullman and P. M. Bentler, *Structural Equation Modeling*, edited by J. A. Schinka and W. F. Velicer, Handbook of Psychology: Research Methods in Psychology Vol. 2 (John Wiley & Sons, Inc., Hoboken, NJ, 2003), pp. 607–634, https://doi.org/10.1002/0471264385.wei0224.

[41] T. A. Brown and M. T. Moore, *Confirmatory Factor Analysis*, edited by R. H. Hoyle, Handbook of Structural Equation Modeling (Guilford Publications, New York, NY, 2012), pp. 361–379.

[42] AERA, APA, and NCME, *Standards for Educational and Psychological Testing* (American Educational Research Association, Washington, DC, 2014).

[43] P. Eaton and S. 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).

[44] J. Han, L. Bao, L. Chen, T. Cai, Y. Pi, S. Zhou, Y. Tu, and K. Koenig, Dividing the force concept inventory into two equivalent half-length tests, Phys. Rev. ST Phys. Educ. Res. **11**, 010112 (2015).

[45] Y. Xiao, G. Xu, J. Han, H. Xiao, J. Xiong, and L. 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] J. Bruun and E. Brewe, 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] M. D. Caballero, E. F. Greco, E. R. Murray, K. R. Bujak, M. Jackson Marr, R. Catrambone, M. A. Kohlmyer, and M. F. Schatz, Comparing large lecture mechanics curricula using the force concept inventory: A five thousand student study, Am. J. Phys. **80**, 638 (2012).

[48] J. M. Nissen, I. H. M. Horses, B. Van Dusen, M. Jariwala, and E. Close, Providing context for identifying effective introductory mechanics courses, Phys. Teach. **60**, 179 (2022).

[49] M. Good, A. Maries, and C. 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] E. Brewe, L. Kramer, and G. O'Brien, Modeling instruction: Positive attitudinal shifts in introductory physics measured with class, Phys. Rev. ST Phys. Educ. Res. **5**, 013102 (2009).

[51] P. Eaton, Evidence of measurement invariance across gender for the force concept inventory, Phys. Rev. Phys. Educ. Res. **17**, 010130 (2021).

[52] T. F. Scott, D. Schumayer, and A. R. Gray, Exploratory factor analysis of a force concept inventory data set, Phys. Rev. ST Phys. Educ. Res. **8**, 020105 (2012).

[53] J. Wang and L. Bao, Analyzing force concept inventory with item response theory, Am. J. Phys. **78**, 1064 (2010).

[54] M. Planinic, L. Ivanjek, and A. Susac, Rasch model based analysis of the force concept inventory, Phys. Rev. ST Phys. Educ. Res. **6**, 010103 (2010).

[55] B. Van Dusen, LASSO: A new tool to support instructors and researchers. American Physics Society Forum on Education Newsletter, 12–14 (2018).

[56] Learning Assistant Alliance, https://learningassistantalliance.org/.

[57] A. Gonzalez-Barrera and M. H. Lopez, Is being Hispanic a matter of race, ethnicity or both? (2015), https://www.pewresearch.org/short-reads/2015/06/15/is-being-hispanic-a-matter-of-race-ethnicity-or-both/ (Retrieved June 2023).

[58] B. O. Muthén, Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes, Psychometrika (to be published).

[59] N. K. Bowen and R. D. Masa, Conducting measurement invariance tests with ordinal data: A guide for social work researchers, J. Soc. Soc. Work Res. **6**, 229 (2015).

[60] G. Hirschfeld and R. von Brachel, Improving multiple-group confirmatory factor analysis in r–a tutorial in measurement invariance with continuous and ordinal indicators, Pract. Assess. Res. Eval. **19**, 7 (2014).

[61] D. Svetina, L. Rutkowski, and D. Rutkowski, Multiple-group invariance with categorical outcomes using updated guidelines: An illustration using m plus and the lavaan/semtools packages, Struct. Equ. Model. **27**, 111 (2020).

[62] R. J. Vadenberg and C. E. Lance, A review and synthesis of the measurement in variance literature: Suggestions, practices, and recommendations for organizational research, Organ. Res. Meth. **3**, 4 (2000).

[63] Y. Rosseel, lavaan: An R package for structural equation modeling, J. Stat. Softw. **48**, 1 (2012).

[64] J. F Hair, Jr., W. C. Black, B. J. Babin, R. E. Anderson, and R. L. Tatham, *Data Analysis Multivariate* (New Jersey. Upper Saddle River. Pearson Education, 2006).

[65] T. D. Jorgensen, S. Pornprasertmanit, A. M. Schoemann, and Y. Rosseel, `semTools`: Useful tools for structural equation modeling (2022), r package version 0.5–6.

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

[67] V. Sawtelle, E. Brewe, and L. H. Kramer, Exploring the relationship between self-efficacy and retention in introductory physics, J. Res. Sci. Teach. **49**, 1096 (2012).

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

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

[70] S. Kanim and X. C. Cid, Demographics of physics education research, Phys. Rev. Phys. Educ. Res. **16**, 020106 (2020).

[71] J. M. Nissen, I. H. M. Horses, B. V. Dusen, M. Jariwala, and E. W. Close, Tools for identifying courses that support development of expertlike physics attitudes, Phys. Rev. Phys. Educ. Res. **17**, 013103 (2021).

[72] P. Barrett, Structural equation modelling: Adjudging model fit, Pers. Individ. Differ. **42**, 815 (2007).

[73] R. B. Kline, *Principles and Practice of Structural Equation Modeling* (Guilford Publications, New York, 2015).

[74] D. Hooper, J. Coughlan, and M. R. Mullen, Structural equation modelling: Guidelines for determining model fit, Electron. J. Bus. Res. Methods **6**, 53 (2008).

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

[76] C. N. McIntosh, Rethinking fit assessment in structural equation modelling: A commentary and elaboration on Barrett (2007), Pers. Individ. Differ. **42**, 859 (2007).

[77] D. A. Kenny and D. B. McCoach, Effect of the number of variables on measures of fit in structural equation modeling, Struct. Equ. Model. **10**, 333 (2003).

[78] B. M. Byrne, *Structural Equation Modeling with LISREL, PRELIS, and SIMPLIS: Basic Concepts, Applications, and Programming* (Psychology Press, New York, 2013).

[79] A. Diamantopoulos, J. A. Siguaw, and J. A. Siguaw, *Introducing LISREL: A Guide for the Uninitiated* (Sage, London, 2000).

[80] Y. Xia and Y. Yang, RMSEA, CFI, and TLI in structural equation modeling with ordered categorical data: The story they tell depends on the estimation methods, Behav. Res. Meth. Instrum. Comput. **51**, 409 (2019).

[81] R. C MacCallum, M. W. Browne, and H. M. Sugawara, Power analysis and determination of sample size for

covariance structure modeling, Psychol. Methods **1**, 130 (1996).

[82] B. G. Tabachnick, L. S. Fidell, and J. B. Ullman, *Using Multivariate Statistics* (Pearson, Boston, MA, 2007), Vol. 5.

[83] P. M. Bentler, Comparative fit indexes in structural models, Psychol. Bull. **107**, 238 (1990).

[84] S. Cangur and I. Ercan, Comparison of model fit indices used in structural equation modeling under multivariate normality, J. Mod. Appl. Stat. Methods **14**, 14 (2015).

[85] J. T Newsom, Structural models for binary repeated measures: Linking modern longitudinal structural equation models to conventional categorical data analysis for matched pairs, Struct. Equ. Model. **24**, 626 (2017).

[86] M. Rhemtulla, P. É. Brosseau-Liard, and V. Savalei, When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions, Psychol. Methods **17**, 354 (2012).

[87] C.-H. Li, Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares, Behav. Res. Meth. Instrum. Comput. **48**, 936 (2016).

[88] G. W. Cheung and R. B. Rensvold, Evaluating goodness-of-fit indexes for testing measurement invariance, Struct. Equ. Model. **9**, 233 (2002).

[89] L. Rutkowski and D. Svetina, Assessing the hypothesis of measurement invariance in the context of large-scale international surveys, Educ. Psychol. Meas. **74**, 31 (2014).

[90] F. F. Chen, Sensitivity of goodness of fit indexes to lack of measurement invariance, Struct. Equ. Model. **14**, 464 (2007).