Bias on the Force Concept Inventory across the intersection of gender and race

John B. Buncher
Department of Physics, North Dakota State University, Fargo, ND, 58108, USA

Jayson M. Nissen
Nissen Education Research and Design, Corvallis, OR, 97333, USA

Ben Van Dusen
School of Education, Iowa State University, Ames, IA, 50011, USA

Robert M. Talbot and Hannah Huvard
School of Education and Human Development, University of Colorado Denver, Denver, CO, 80217, USA

Education researchers often compare performance across race and gender on research-based assessments of physics knowledge to investigate the impacts of racism and sexism on physics student learning. These investigations' claims rely on research-based assessments providing reliable, unbiased measures of student knowledge across social identity groups. We used classical test theory and differential item functioning (DIF) analysis to examine whether the items on the Force Concept Inventory (FCI) provided unbiased data across social identifiers for race, gender, and their intersections. The data was accessed through the Learning About STEM Student Outcomes platform and included responses from 4,848 students posttests in 152 calculus-based introductory physics courses from 16 institutions. The results indicated that the majority of items (22) on the FCI were biased towards a group. These results point to the need for instrument validation to account for item bias and the identification or development of fair research-based assessments.

2021 PERC Proceedings edited by Bennett, Frank, and Vieyra; Peer-reviewed, doi.org/10.1119/perc.2021.pr.Buncher

I. INTRODUCTION

Social movements across the world have drawn an increased attention on the role that racism and sexism play in creating unjust social systems and outcomes [1]. Education researchers often compare performance across race and gender on research-based assessments of physics knowledge [2, 3] to investigate the impacts of racism and sexism on physics student learning. These investigations' claims rely on research-based assessments providing reliable, unbiased measures of student knowledge across social identity groups. The development of research-based assessments usually relies on data from students at the institution from which the instrument was developed, which are often research intensive, highly selective institutions with physics courses that over-represent White [4] and Asian men. The common lack of diversity in these studies [5] limits the generalizability of their validity arguments.

In this investigation, we used classical test theory [6] and differential item functioning (DIF) [7] analysis to examine whether the items on the Force Concept Inventory (FCI) [8] provided unbiased data across social identifiers for race and gender. Prior investigations of the FCI have found evidence of item bias across genders [9, 10]. Our investigation expands on these findings by performing an intersectional analysis that examines the potential for bias across genders and races.

II. RESEARCH QUESTION

To better understand the potential for items on the FCI to bias student performance data across social identity groups for gender and race, we asked the following questions.

1. Which items on the FCI, if any, show bias in student performance across social identity groups?
2. What trends exist, if any, in the item biases across social identity groups?

III. BACKGROUND

A. Instrument Validation

The research questions we pose are central to instrument validity. If test items are biased for or against a group or subgroup of respondents, then the target construct is being measured differentially for those groups or subgroups. Therefore any inferences drawn are likely to be confounded due to this bias. Contemporary conceptions of validity focus on building an argument for validity [11, 12]. Instrument validity is not a binary condition, and instruments are not "validated." Our work in this paper constitutes only one part of a validity argument for the FCI. While many questions have been raised about the construct validity of the FCI [13, 14], we are not aware of any work on FCI validity which considers the items themselves in terms of response processes, and potential for differential item functioning across respondents by race or gender except for the work across gender by Traxler *et al.* [9] and Henderson *et al.* [10].

B. Differential Item Function Analysis

Differential Item Functioning (DIF) is a general term for describing how test items may perform differently or unexpectedly for subgroups responding to those items [15, 16]. In our work, those subgroups are defined by social identifiers for gender and race. By comparing item scores for social identifier subgroups having the same overall FCI scores, we can see if the items are functioning differentially for certain subgroups. Any indication of DIF may be an indicator of item bias, either for or against a particular subgroup.

Many statistical methods can examine DIF. In this work, we use the Mantel-Haenszel method [7]. This method uses chi-squared contingency tables to compare item scores between subgroups for all items on a test. We used White men as the reference group and minoritized groups as the focal groups. The range of total scores on the test are divided into intervals that then serve as the basis for matching members of the subgroups. Contingency tables for each interval are then constructed to compare subgroups on item performance. Differences between scores on an item for students with similar ability levels (i.e., overall exam score) across subgroups (e.g., race and gender) provide evidence of item bias.

C. Existing Literature

Previous validity work related to the FCI has focused on factor analysis, which has produced inconsistent results [13]. Little work has investigated potential biases of the items themselves, an important component of validity. Traxler *et al.* [9] were the first to publish item-level analysis of the FCI using Differential Item Functioning (DIF). The authors examined how specific items were biased across gender and concluded that items 12, 14, 21, 22, 23, and 27 were biased in favor of men. Additionally, items 9 and 15 were identified to be biased in favor of women.

Based on their results and previous research, the authors recommended a refined 19-item FCI that does not include their identified biased items nor items that demonstrated poor reliability (items removed were 6, 9, 12, 14, 15, 21, 22, 23, 24, 27, and 29). The authors explain, "Because the FCI has not demonstrated a consistent factor structure and therefore is primarily a single factor instrument measuring the degree to which a student possesses a 'Newtonian force concept,' a 19-item instrument should measure this construct with approximately the same accuracy as a 30-item instrument." [9]

The assumption that the 19-item FCI would measure the same construct as the original 30-item FCI has large implications for validity. Several measurement methodologists argue

that instrument functioning and validity should be established each time a psychometric instrument is edited from its original form or used within a new context [12, 17]. Thus, Traxler *et al.'s* refined 19-item FCI should undergo the same rigorous testing as the original 30-item instrument to determine how it is functioning across various demographic groups (not just gender). This is especially important considering that the techniques used in Traxler *et al.* [9] were sample variant, so their findings may not hold when the FCI is used in a different population or context.

Additional studies have examined potential bias in the FCI, but have focused on differences of overall scores, not individual item bias [18, 19]. Planinic *et al.* [20] used a Rasch model to examine the dimensionality of the FCI (which they found to be sufficiently unidimensional), but did not perform a DIF analysis to assess potential biases across any social identity groups. Considering that Traxler *et al.* [9] did find several items that were problematic in terms of their biases toward gender, there exists a need to examine how the FCI may be biased across several social identity groups beyond gender.

IV. CONCEPTUAL FRAMEWORK

A. Quantitative Critical Race Theory (QuantCrit)

We used a Quantitative Critical (QuantCrit) framework [21, 22] in this investigation. Below, we describe four principles of QuantCrit and the ways we strove to embody them:

1. *The centrality of oppression* - We assumed that racism and sexism are present throughout society that we must explicitly examine lest our statistical models legitimize existing inequities. Scientists often view science as having a culture of no culture [23–25] which could lead them to incorrectly assume that items on an instrument will perform the same across all groups. In the case of the FCI, we assume that its validation work was performed disproportionately with White men as little evidence supports its validity for minoritized groups.
2. *Categories are neither 'natural' nor given* - All data are socially constructed and reflect the hegemonic power structures that created them. Our analyses aggregated students by race and gender. These categories do not represent any natural or scientific truth about students but are social constructs that maintain hegemonic power structures. The dynamic socially-negotiated natures of race and gender does not diminish the very real effects of racism and sexism associated with them. We strive to represent student self-identified genders and races with as much fidelity our data will allow. For example, a meaningful number of students who identified as Hispanic also identified as White. Because we had sufficiently large sub-group sample sizes to do so, we reflected these distinctions in students' identities in our analysis by modeling Hispanic, White Hispanic, and White non-Hispanic as three distinct groups.
3. *Data is not neutral and cannot 'speak for itself'* - Racist and sexist assumptions can shape every stage of collecting, analyzing, and interpreting data [26]. Our investigation is focused on how items on the FCI may bias data in often unexamined ways. In analyzing the data, we strove to examine potential biases that could arise from our methods. For example, we broke from the traditional practice of only including effect sizes that were statistically significant. P-values depend on sample sizes and can lead researchers to dismiss meaningful inequities due to lack of representation in minoritized groups [27]. Instead, we focus on how meaningful the differences in item performance are between groups. We also used the consistency of results across groups to inform our level of certainty about an item's bias.
4. *The importance of intersectionality* - Identity is multifaceted (e.g., race, gender); each aspect dynamically intersects with each other and society's associated oppressive power structures to shape experience [28]. In this analysis, we accounted for the dynamic interactions between sexism and racism by examining each combination of genders and races separately.

V. METHODS

The data came from the Force Concept Inventory (FCI) [29]. We accessed the data through the Learning About STEM Student Outcomes (LASSO) platform's [30] research database. The LASSO platform collects large-scale, multi-institution data by administering, scoring, and analyzing research-based assessments online. The research database only includes anonymized data for students who consented to share their data with researchers. The data came from 4,848 students posttests administered at the end of 152 calculus-based introductory physics courses from 16 institutions.

To clean the data, we removed the score if the student took less than 5 minutes or answered less than 80% of the items. We then removed courses with less than 10 students or less than 40% participation on the pretest or posttest.

We analyzed the data by comparing subsets of students based on their social identifiers to the White men students whom we reasoned were the most benefited and least harmed by White supremacy and patriarchy. The data set included social identifier data for gender and race. We only investigated scores for populations with at least 20 students total [31]. This guideline precluded investigating bias for transgender, Hawaiian or Pacific Islander, or Native American or Alaskan Native students. The social identity groups in the study included men and women for gender and Asian, Black, non-White Hispanic, White non-Hispanic, and White Hispanic for race. The non-White Hispanic group includes all students that identified as Hispanic and did not identify as White, most of whom chose 'a race not listed' or did not choose a race. To simplify discussion we will refer to non-White Hispanic students as "Hispanic" and White non-Hispanic stu-

dents as "White". Table I shows the descriptive statistics for the posttest scores for these groups of students.

TABLE I: Descriptive statistics by gender and race.

	Women			Men		
Race	N	Mean	St.Dev	N	Mean	St.Dev
Asian	238	62.2	23.0	364	64.2	24.5
Black	119	47.3	20.9	128	49.7	21.1
Hispanic	90	48.6	24.9	253	54.0	20.8
White Hispanic	115	52.4	21.8	273	61.6	23.2
White	784	61.4	23.1	2206	71.3	21.1

In classifying the amount of DIF for each item, we used the Educational Testing Service (ETS) classification scale [32]. This scale transforms the odds ratio α_{MH}, such that items that are equally likely to be answered correctly by both groups ($\alpha_{MH} = 1$) correspond to $\Delta\alpha_{MH} = 0$. The thresholds for whether items have a "Large" or "Moderate" effect size are where the reference group (i.e., White men) is more than 50% and 90% more likely to answer correctly than the focal group (i.e., a minoritized group; $\alpha_{MH} = 1.53$, $\Delta\alpha_{MH} = 1$ and $\alpha_{MH} = 1.9$, $\Delta\alpha_{MH} = 1.5$, respectively). Items with $|\Delta\alpha_{MH}| < 1$ having a statistically significant χ^2 are classified as having a "small" effect size, we do not list these.

It is important to note that the Mantel-Haenszel test characterizes items as DIF items when an item functions differently than the overall instrument. If the reference group (White men in this study) performs better on the instrument overall, items detected as favoring the focal group may just be favoring the reference group to a smaller degree.

VI. FINDINGS

White men scored higher than all other social identifier groups on nearly every item on the FCI. Figure 1 plots the classical test theory (CTT) difficulty for the reference group (i.e., White men) on the vertical axis and the focal group (i.e., Asian men or Black women) on the horizontal. Due to space limitations Fig. 1 only includes two of the nine plots. We chose these two plots because they represented the groups with the least spread (Asian men) and most spread (Black women) in item difficulties. CTT difficulty is the proportion of students in that group who chose the correct answer. Items with equal performance across groups lie close to the diagonal. As shown in Fig. 1, nearly every point falling above the diagonal indicated that White men scored higher than the focal groups. The exceptions to this were White Hispanic men on items 17 and 29; White women on item 29; Black men on item 29; and Black women on items 4, 15, and 29.

Before reporting DIF results, we reiterate that the Mantel-Haenszel test detects items that behave differently than the overall instrument. Since White men scored higher on nearly every item compared to every other group, White men often perform better even on items that "favored" the focal groups, such as item 26 in Fig. 1. Table II shows the items with large and moderate effect sizes in favor of White men or one of the nine focal groups. We list all items identified as having "large" or "moderate" DIF effect size and list whether they favor the reference group (White men) or the focal group. We also indicate which of these items are statistically significant. We do not list any items with "small" DIF.

As discussed in the Conceptual Framework, we focused on the magnitude of the DIF scores, rather than p-values. The DIF analyses identified more items as having a large or moderate DIF (favoring either group) when the focal group was women (58 items) than minoritized men (25 items). Ten items had large or moderate bias favoring White men compared to minoritized women and five of these items also favored White men compared to minoritized men.

DIF identified several items as favoring White men across multiple groups. Item 14 favored White men over all groups of women. Item 27 favored White men over Asian, Black, and Hispanic women. Item 23 favored White men over every group other than White Hispanic men and White women. Item 10 favored White men over every group other than Asian men, White non-Hispanic women, and White Hispanic men. No items favored White men over White Hispanic men.

Several items favored minoritized students. Item 4 favored every group over White men other than White Hispanic men. Item 29 favored every group over White men other than Asian men and women. Item 28 favored every group over White men other than Black women and White Hispanic men. Item 9 seems to favor women of color, showing moderate-to-large DIF favoring Asian, Black, and Hispanic women, but not for any group of men.

VII. CONCLUSIONS

Our results add to the growing evidence that the FCI functions differentially for students across social identifiers. It is possible, but unlikely, that random divisions would find differences across groups. The fact that we found so many meaningful DIF scores, some of which were consistent across multiple groups, provides a strong indication that items were acting differently across social identifiers. Items 10, 14, 23, and 27 appear especially problematic, favoring White men over most other groups to a greater degree than the FCI as a whole. These results support what Traxler *et al.* [9] found with items 14, 23, and 27 as having a large DIF favoring men. Traxler *et al.* also found items 12, 21, and 22 as favoring men. While we did find these three items as having medium-to-large DIF for some groups, we do not see any clear pattern to the groups they favored.

Items that favored minoritized groups more than the FCI as a whole were also consistent with Traxler *et al.* Specifically, item 4 showed moderate-to-large DIF for 8 of our 9 groups, item 28 for 7 of the 9 groups, and item 29 for 7 of 9 groups.

Removal of items that are problematic may reduce the performance differences seen on the FCI, but it introduces two

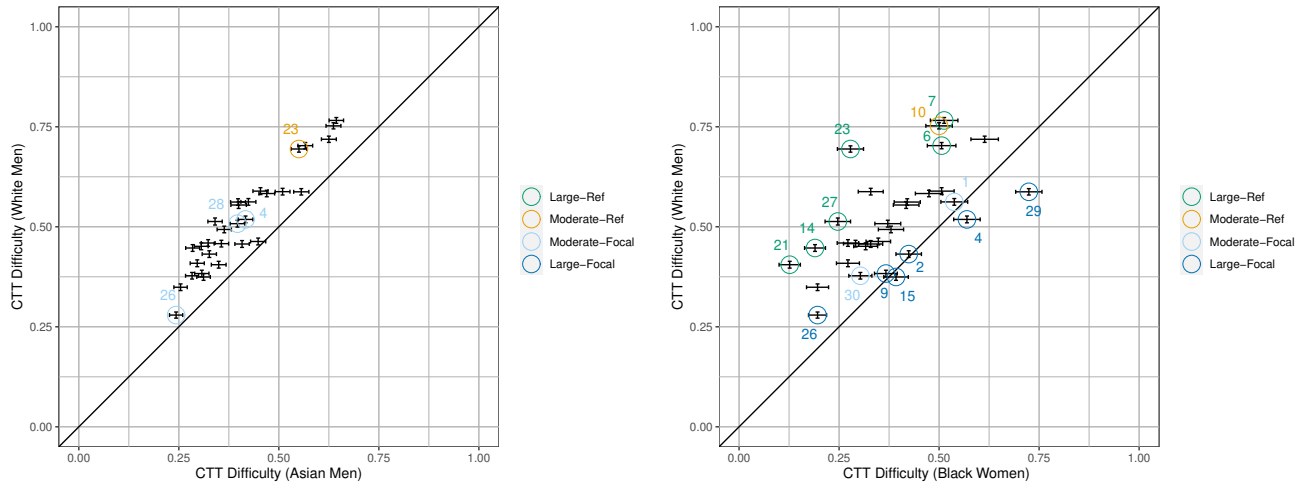

FIG. 1: FCI Posttest CTT DIF results. CTT difficulty is the percentage correctly answering an item. Items flagged as having moderate-to-large DIF scores are circled. Of the nine comparisons made, the figure for Asian men had the least spread and the the figure for Black women had the largest spread.

TABLE II: Items with moderate or large effect sizes that favored the reference group (White men), or a focal group (minoritized student). * indicates items that were statistically significant at the $p < 0.05$ level after Bonferroni correction.

Focal Group	Favors White men		Favors Focal Group	
	Large	Moderate	Moderate	Large
Asian men		23*	4, 26*, 28	
Black men	21*, 23*	7, 10*, 14	1, 2, 4, 15, 28, 29*	30*
Hispanic men	23*	7*, 10*	28, 29*	4*
White Hispanic men			25*, 29*	
Asian women	14*, 27*	6, 7*, 10*, 12, 22*, 23*, 24	18*, 26	4*, 9*, 25*, 28*
Black women	6*, 7*, 14*, 21*, 23*, 27*	10*	1, 30*	2*, 4*, 9*, 15*, 26*, 29*
Hispanic women	7*, 10*, 14, 21*, 23*, 27*		1, 2, 30	4*, 9, 15*, 26, 28*, 29*
White Hispanic women	10*, 14*, 23*	6, 12, 21	5*, 15	4*, 25*, 26, 28, 29*
White women	14*	23*	4*, 28*, 29*	

problems. First, removing an item from an instrument breaks its validation argument requiring that the new instrument be re-validated to ensure that intended construct is being measured. Second, DIF analysis identified 22 items as biased. Removing them would only leave 8 items, further bringing into question the ability of the instrument to measure the construct of interest. Ideally, this analysis would be completed during the development of an instrument, rather than as a post-hoc analysis. To truly address these issues would require the creation of a new instrument.

VIII. LIMITATIONS AND FUTURE WORK

In future work we will investigate these methods using Item Response Theory (IRT). IRT will provide information about how the items perform across different student ability levels [33]. This information can tell us if certain items favor students with high (or low) abilities, which could manifest as gender or racial biases due to inequities in students' current and prior educations. Because IRT is invariant across test populations and handles differences in group variances better than CTT, IRT may assist us in understanding the patterns in the bias identified by DIF analysis.

Aggregations across social identifier groups may hide further biases in the FCI. Including socioeconomic status and further disaggregating the data may reveal additional biases on physics assessments.

Future research will investigate other first and second semester physics research-based assessments.

IX. ACKNOWLEDGEMENTS

We wish to thank NSF for project funding (DUE-1928596).

[1] C. J. Lebron, *The making of black lives matter: A brief history of an idea* (Oxford University Press, 2017).

[2] S. Singer and K. A. Smith, Discipline-based education research: Understanding and improving learning in undergraduate science and engineering, Journal of Engineering Education **102**, 468 (2013).

[3] J. L. Docktor and J. P. Mestre, Synthesis of discipline-based education research in physics, Physical Review Special Topics-Physics Education Research **10**, 020119 (2014).

[4] In this publication, we capitalize all races, including White, emphasizing that there is no default race and that they are all social constructs with associated sets of cultural practices.

[5] S. Kanim and X. C. Cid, Demographics of physics education research, Physical Review Physics Education Research **16**, 020106 (2020).

[6] L. Crocker and J. Algina, *Introduction to classical and modern test theory.* (ERIC, 1986).

[7] N. Mantel and W. Haenszel, Statistical aspects of the analysis of data from retrospective studies of disease, J. Natl. Cancer Inst. **22**, 719 (1959).

[8] D. Hestenes, M. Wells, and G. Swackhamer, Force concept inventory, The Physics Teacher **30**, 141 (1992).

[9] A. Traxler, R. Henderson, J. Stewart, G. Stewart, A. Papak, and R. Lindell, Gender fairness within the force concept inventory, Physical Review Physics Education Research **14**, 010103 (2018).

[10] R. Henderson, J. Stewart, and A. Traxler, Partitioning the gender gap in physics conceptual inventories: Force concept inventory, force and motion conceptual evaluation, and conceptual survey of electricity and magnetism, Physical Review Physics Education Research **15**, 010131 (2019).

[11] 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).

[12] M. T. Kane, Validation, in *Educational Measurement*, edited by R. L. Brennan (American Council on Education and Praeger, 2006) pp. 17–64, 4th ed.

[13] D. Huffman and P. Heller, What does the force concept inventory actually measure?, Phys. Teach. **33**, 138 (1995).

[14] M. M. Hull, J.-I. Yasuda, M.-A. Taniguchi, and N. Mae, Towards quantification of the FCI's validity: the effect of false positives, in *2017 Physics Education Research Conference Proceedings* (2018) pp. 180–183.

[15] G. Camilli, Test fairness, in *Educational measurement*, edited by R. L. Brennan (American Council on Education, Westport, CT) pp. 220–256.

[16] B. 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).

[17] R. Komperda, K. N. Hosbein, and J. Barbera, Evaluation of the influence of wording changes and course type on motivation instrument functioning in chemistry, Chem. Educ. Res. Pract. **19**, 184 (2018).

[18] J. Docktor and K. Heller, Gender differences in both force concept inventory and introductory physics performance, AIP Conf. Proc. **1064**, 15 (2008).

[19] M. Mears, Gender differences in the force concept inventory for different educational levels in the united kingdom, Phys. Rev. Phys. Educ. Res. **15**, 020135 (2019).

[20] 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).

[21] F. K. Stage, Answering critical questions using quantitative data, New Directions for Institutional Research **2007**, 5 (2007).

[22] D. Gillborn, P. Warmington, and S. Demack, Quantcrit: education, policy, 'big data' and principles for a critical race theory of statistics, Race Ethnicity and Education **21**, 158 (2018).

[23] J. S. Taylor, Confronting "culture" in medicine's "culture of no culture", Academic Medicine **78**, 555 (2003).

[24] B. Subramaniam and M. Wyer, Assimilating the "culture of no culture" in science: Feminist interventions in (de) mentoring graduate women, Feminist Teacher , 12 (1998).

[25] S. Traweek, *Beamtimes and lifetimes* (Harvard University Press, 2009).

[26] A. Covarrubias, P. E. Nava, A. Lara, R. Burciaga, V. N. Vélez, and D. G. Solorzano, Critical race quantitative intersections: A testimonio analysis, Race Ethnicity and Education **21**, 253 (2018).

[27] V. Amrhein, D. Trafimow, and S. Greenland, Inferential statistics as descriptive statistics: There is no replication crisis if we don't expect replication, The American Statistician **73**, 262 (2019).

[28] K. Crenshaw, Mapping the margins: Intersectionality, identity politics, and violence against women of color, Standard Legal Review **43**, 1241 (1990).

[29] D. Hestenes, M. Wells, and G. Swackhamer, Force concept inventory, The physics teacher **30**, 141 (1992).

[30] B. Van Dusen, Lasso: A new tool to support instructors and researchers, American Physics Society Forum on Education Fall 2018 Newsletter (2018).

[31] J. P. Simmons, L. D. Nelson, and U. Simonsohn, False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant, Psychological science **22**, 1359 (2011).

[32] N. J. Dorans and P. W. Holland, *DIF detection and description: Mantel-Haenszel and standardization*, Tech. Rep. (1992).

[33] "Ability level" is the term that IRT uses to classify the amount of a construct that a student is measured to have. However, it is a potentially problematic term in the context of equity research as it could support deficit narratives. We will examine alternative terms as we engage more fully in this work.