Mediational effect of prior preparation on performance differences of students underrepresented in physics

John Stewart,[1,*] Geraldine L. Cochran,[2] Rachel Henderson,[3] Cabot Zabriskie,[1] Seth DeVore,[1] Paul Miller,[1] Gay Stewart,[1] and Lynnette Michaluk[4]

[1]*Department of Physics and Astronomy, West Virginia University, Morgantown, West Virginia 26506, USA*

[2]*Department of Physics and Astronomy, Rutgers, The State University of New Jersey, Piscataway, New Jersey 08854, USA*

[3]*Michigan State University, Department of Physics and Astronomy, East Lansing, Michigan 48824, USA*

[4]*West Virginia University Center for Excellence in STEM Education, Morgantown, West Virginia 26506, USA*

(Received 23 June 2020; accepted 25 January 2021; published 10 February 2021)

This study examined the mediation and moderation of membership in a demographic group underrepresented in physics classes on course outcomes measured by course grades and Force and Motion Conceptual Evaluation (FMCE) post-test scores. The study used a large dataset ($N = 4490$) of course grades, SAT and ACT mathematics scores (ACTM), and matched FMCE pretest and post-test scores to investigate differences by gender, underrepresented ethnic or racial minority (UERM) status, and status as a first-generation college student (FGCS). For UERM and FGCS students, ACTM and pretest scores significantly mediated the relation of membership in the demographic group and both course grade and post-test score. Differences between minority and majority members of these groups were largely removed by controlling for ACTM and pretest scores. The overwhelming majority of the effect acted through ACTM for course grade (60% and 45%, respectively), while more of the effect acted through pretest score for the post-test (36% and 48%, respectively). As such, for these groups prior preparation measures predict physics outcomes (course grades or post-test scores) differently. The mediational relations for gender were dramatically different. No mediation was detected for the relation of gender to course grade because no significant difference in course grade existed. Sixty percent of the effect of gender on post-test score was not explained by either ACTM or pretest score; pretest score accounted for 38% of the effect. As such, the majority of the difference in post-test scores between men and women was not explained by either ACTM or pretest scores. Significant moderation was also detected showing that the relation of these variables was not consistent for members of all demographic groups.

DOI: 10.1103/PhysRevPhysEducRes.17.010107

I. INTRODUCTION

Students come to physics classes with differing levels of general academic preparation and specific prior preparation in physics. These differences may influence the overall course outcome of the students measured by course grades or the student's conceptual learning measured by conceptual inventory post-test scores. Prior academic preparation may not be the same for all demographic subgroups represented in physics classes. The reasons for these differences are disparate. For example, women elect high school physics courses at a lower rate than men (while electing more advanced high school chemistry classes) [1,2] and underrepresented minority students may have less access to high school physics or advanced high school mathematics classes [3]. Differences in preparation by students underrepresented in physics may lead to varying outcomes and further contribute to the underrepresentation of these students. Understanding whether differences in preparation exist and how they contribute to differences in physics performance may be one key to promoting equitable learning experiences for all physics students.

A. Research questions

This study explores the relation of membership in a demographic subgroup underrepresented in physics classes, the academic preparation of that group, and the course outcomes of that group. It further seeks to determine

*jcstewart1@mail.wvu.edu

if the relationship between these variables is consistent between the minority and majority members of the demographic group and if there are any additional effects from being a member of multiple underrepresented groups.

This study seeks to answer the following research questions:

RQ1: Are there differences in conceptual post-test scores and course grades between demographic groups in physics classes?

RQ2: How are these achievement differences related to general academic preparation and prior preparation in physics?

RQ3: Are these relations consistent for the minority and majority members of the demographic groups?

RQ4: Are there additional effects of being a member of multiple underrepresented groups?

B. Positionality

The motivation of this work was to extend work being performed to understand gender differences in conceptual learning assessed by research-based multiple choice instruments to other demographic groups underrepresented in physics classes to ensure equitable treatment of all students. Research that identifies performance gaps without acknowledging the historical and current inequities and injustice that led to these differences in performance exacerbates these inequities and injustices. Thus, we acknowledge that students from minoritized and marginalized ethnic or racial groups, particularly Black/African-American, Hawaiian/Pacific Islander, and Latinx students, have been disadvantaged in the educational system at many levels. Furthermore, systemic racism impacting various aspects of their lives also influences performance in educational settings.

Various studies have used the terms Hispanic, Latinx, or Hispanic/Latinx to refer to people identifying as being of Latin American ethnic identity, native speakers of Spanish, or having Spanish-speaking ancestry, we will use the term Latinx throughout this paper.

The authorial team is comprised of seven physics education researchers and one social science researcher. The team includes instructors with over 50 years of combined experience teaching introductory physics in large lecture settings. The authorial team includes four members identifying as women, four as men, seven as White, one as Black, and three were first-generation college students. While we acknowledge that investigations of differences in group performance are deeply connected to inequity and injustice, we think it is important to identify such gaps. Previous research—see Sec. I E—reveals that performance gaps (based on race or gender and race) are apparent in early educational experiences [4–6]. Our study investigates whether or not these gaps persist in the university physics environment. In addition, Parks and Schmeichel call for researchers to start "placing the burden for discussing race and racism on educational communities as a whole rather than solely on the shoulders of those scholars whose work primarily deals with equity and social justice" [7]. Furthermore, given that this is one of the few studies to look at the interactions of race, gender, and first-generation status, it provides a baseline to see if efforts to mitigate the disadvantages and inequities impacting minoritized and marginalized students have resulted in changes in performance for these groups in comparison to themselves; thus, this study provides a baseline for future equity of individuality studies [8].

C. Equity

The term equity is used in a variety of ways in the literature. Rodriguez *et al.* [8] suggest clarifying the way that equity is used in research and being cognizant of the shortfalls of using various equity models as they impact the design of the research and the interpretation of the findings. This study utilizes an equity of parity model, which defines equity as mitigating differences in students' prior preparation by reducing the differences on post-instruction performance measures such as post-test scores or course grades. For additional discussion of this and other equity models see Rodriguez *et al.* [8].

D. Intersectionality

This work also begins the much needed move to a more complete description of student identity which intersectionality theory suggests is necessary to understand the dynamics of discrimination and privilege [9]. Intersectionality theory posits that multiple social identities intersect and interact to produce patterns of privilege or disadvantage. The intersection of various sociodemographic identities defined in terms of relative power and privilege may result in unique experiences for individuals [10]. This work, which examines the intersection of gender, underrepresented ethnic or racial minority status, and first-generation college student status, is only a beginning; many other intersecting identities interact and may be important for understanding the students studied: the dynamics of the Appalachian culture, socioeconomic status (SES), the role of an urban or rural upbringing, and veteran status to name only a few. Substantially more research is needed to fully understand how the interactions of these identities influence the student's experiences and treatment in physics classes—and outside of physics classes—and what additional supports can provide each student an equal opportunity to succeed.

A few studies have investigated the effects of intersecting identities in physics. Rosa and Mensah [11] found that African-American women faced unique challenges while pursuing graduate education in physics. Hazari *et al.* [12] found that Latinas faced the greatest challenges in science identity development, including physics identity development. Ong also investigated the intersection of race and gender for women of color in physics and highlighted

their unique experiences [13]. Ko, Kachchaf, Hodari, and Ong investigated the unique navigational strategies used by women of color in physics to persist and remain successful [14].

E. Results of prior studies

This work seeks to answer the call of Scherr [15] to extend work on inclusion in physics beyond gender. This study also seeks to extend quantitative analysis to groups other than White male students who have been overstudied in physics education research (PER) [16] within the conceptual framework of mediation and moderation which is discussed in detail in Sec. II C.

This work defines underrepresented ethnic or racial minority (UERM) students as students who identify as Black/African-American, Latinx, American Indian/Alaska Native, or Hawaiian/Pacific Islander. This classification has been used as "underrepresented minority" (URM) in prior studies including Salehi *et al.* described below.

This work extends the study of Salehi *et al.* [17] examining the effect of general academic preparation, operationalized by Scholastic Aptitude Test (SAT) mathematics percentile score (SATM), and prior preparation in physics, operationalized by physics pretest score, on final examination scores in introductory physics. Salehi *et al.* showed that significant differences in final exam performance were observed between women, UERM students, first-generation college students (FGCS), and students who are not members of these groups (men, White non-Latinx students, and non-first-generation college students, respectively). These differences ranged from small to medium effects using Cohen's d criteria (small effect $d = 0.2$, medium effect 0.5, and large effect 0.8 [18]). These differences were no longer significant when SATM and pretest score were controlled for by adding these variables to the regression equations. Salehi *et al.* concluded with a call to replicate this work at other institutions. The current work was partially developed as a response to this call.

There were features of the three samples ($N = 2669$, 786, and 378) used in Salehi *et al.* which suggest the results may not be fully generalizable. The samples were all composed of very well prepared students with average SATM of 89%, 97%, and 97% and average pretest scores [either Force Concept Inventory (FCI) [19] or Force and Motion Conceptual Evaluation (FMCE) [20]] of 43.5%, 55.5%, and 62% (the average of the ends of the ranges reported). This raises the concern that the results may only apply to the most thoroughly prepared students. Kamin and Cid caution that historically PER has examined students who are more mathematically prepared than the general population of physics students [16]. In contrast, the students in the present study have an average ACT mathematics percentile score of 80% and FMCE pretest percentile score of 23%. Two of the samples in Salehi *et al.* may have been insufficiently large to investigate interaction effects. The current study analyzes a substantially larger sample.

Final exam grades or course grades ultimately have a substantial impact on students' college outcomes. In a multi-institution study, Hazari *et al.* [6] investigated the role of gender and UERM status on college physics course grade. Women, Latinx students, and African-American students earned significantly lower course grades than male White non-Latinx students controlling for SATM and high school mathematics preparation. While pretest data were not presented, the study controlled for a number of features of high school physics classes which should affect pretest scores. Controlling for these features changed the regression coefficients for Latinx and African-American students little while the regression coefficient for women increased in magnitude (but was no longer significant). This study also investigated interactions and found some features of high school preparation that affected men and women differently.

In PER, a significant strand of research has investigated students' conceptual learning and differences in that learning by demographic group. Much of the research within PER has focused on the "gender gap" in physics conceptual inventory scores. In a review of many studies, Madsen, McKagan, and Sayre [21] reported that on conceptual mechanics evaluations, the FCI, and the FMCE, men score 13% higher than women on the pretest and 12% on the post-test.

Kost, Pollock, and Finkelstein showed that FMCE post-test differences between men and women were explained by a combination of FMCE pretest, a composite mathematics preparation score, and the student's attitudes about science [22]. They also report a significant gender-by-pretest interaction. Ethnicity was not a significant predictor of post-test performance on the FMCE controlling for pretest score; however, they acknowledge that the low number of UERM students in their sample was probably insufficient to resolve the effect of race and/or ethnicity.

Henderson, Stewart, and Traxler [23] explored gender differences in post-test scores using methods similar to those in Salehi *et al.* [17]. Unlike in Salehi *et al.*, which found differences in the final exam average became insignificant when controlling for SATM and pretest scores, substantial differences in post-test performance remained after controlling for these variables.

Brewe *et al.* [24] reported significant differences in FCI post-test scores between men and women and between UERM and White non-Latinx students. The differences between UERM and majority students were no longer significant controlling for SATM; however, the differences between men and women remained significant controlling for SATM.

Henderson and Stewart examined differences in the FMCE post-test scores and physics course grades of underrepresented students [25] finding significant differences in course grade between different demographic groups ranging

up to a medium effect and post-test differences ranging up to a large effect.

Henderson, Zabriskie, and Stewart [26] reported the effect sizes of post-test performance differences on the FMCE by gender ($d = 0.48$), UERM status (0.24), and FGCS status (0.12). Controlling for SAT or ACT mathematics scores (ACTM) substantially reduced the regression coefficients of UERM and FGCS students; for FGCS students the coefficient was no longer significant. For UERM students, while the coefficient was still significant, it represented a small effect size. The regression coefficient for women was virtually unchanged by controlling for ACTM. Both Henderson and Stewart [25] and Henderson, Zabriskie, and Stewart [26] drew their sample from the same student population as the current work.

F. Minoritized ethnic or racial groups

In 2016, the U.S. Census Bureau reported that African Americans made up 13.3% of the U.S. population and Latinx 17.8% [27]. African-American and Latinx students enter college at a slightly lower rate than other students. African-American and Latinx students represented 27% (12.6% African American and 14.4% Latinx) of first-term students at public 4-year institutions [28]. African-American and Latinx students are underrepresented in many science, technology, engineering, and mathematics (STEM) majors. For example, in engineering and engineering technology, 6.7% of undergraduate majors are African American and 12.7% are Latinx [29].

1. Ethnicity or race and performance in high school

The underrepresentation of students by race or ethnicity in STEM courses is not limited to the college environment. The *National Science Board: Science and Engineering Indicators 2016* report showed that by 11th grade 14.8% of African-American and 17.1% of Latinx students enrolled in advanced science courses (Physics 2, Chemistry 2, Biology 2), far fewer than their White, 21.7%, and Asian, 37.3%, peers [5]. For many African-American and Latinx students, the lack of advanced math and science courses taken in high school pushes them out of the STEM pipeline, as advanced course taking is vital to STEM degree success [30]. Compounding this issue is the lack of advanced courses offered in high schools in communities or districts that serve African-American and Latinx students [3]. Differences in class-taking patterns are reflected in differences in standardized test scores. According to the *ACT Profile Report 2016*, African-American students had an average ACT mathematics score of 17.0, Latinx students 18.8, White students 21.7, and Asian students 25.0 [4]. The underrepresentation of African-American and Latinx students in high school science classes may lead to preparation differences that cause performance differences in college physics classes.

2. Ethicity or race and performance in college

The National Center for Education Statistics reported that 41% of African-American and 53% of Latinx students enrolled in 4-year institutions complete a degree in 6 years at that institution while 63% of White and 71% of Asian students complete a degree in 6 years [31]. White women, women of color, and men of color are underrepresented among STEM degree holders in part because they leave STEM majors at substantially higher rates than White or Asian male students [32–34].

Toven-Lindsey *et al.* [34] reported that UERM students entering colleges in the US are just as likely as their White peers to aspire to complete a STEM major [35–37]. A study by Riegle-Crumb and King suggested that controlling for prior preparation explained the difference in representation of White and UERM students [38]. Upon further investigation, they found that controlling for high school preparation, male African-American students were more likely than White male students to elect a physical science or engineering major and African-American women have smaller major selection differences than White women [39]. Trusty, Ng, and Plata suggested that, while there is an effect of race on initial major selection, this effect can be substantially diminished, especially among women, for students of high SES [40].

G. Gender

Gender differences between male and female students have been reported on multiple standardized examinations including the SAT [41] and the Graduate Record Examination (GRE) [42]. Differences in academic performance measured by course grade have also been reported with women consistently achieving higher course grades than men [43].

Extensive research has been performed examining differences in the performance and representation of men and women in STEM. Men and women elect STEM classes at different rates in high school. Men take physics at a 5.6% higher rate than women, while women take more advanced chemistry and biology courses [1,2]. Differences are also measured in the ACT College Readiness STEM benchmark [44] with 30% of men and 21% of women meeting the benchmark. Research has shown that having a high school physics class is related to higher college physics grades [6,45].

Many explanations have been advanced to explain gender differences in general academic achievement and achievement in physics including cognitive differences [46–50], sociopsychological factors including math anxiety [51,52], science anxiety [53–55], and stereotype threat [56–59], instrumental bias in physics conceptual instruments [60–64], and the method of physics instruction [24,65–70]. For a thorough review of gender differences in academic achievement see Henderson *et al.* [71].

This study reports gender as a binary variable; this reporting of gender is consistent with other PER studies of gender differences. We acknowledge that this is not optimal and, as more complete data are collected, a more complex treatment of gender should be performed as suggested by Traxler *et al.* [72].

H. Ethnicity or race and gender

Very little quantitative work has explored the combined effects of UERM status and gender on STEM academic performance. In a study of 9813 middle school students from 10 U.S. states, Scafidi and Bui found no moderation effect of gender or race for middle school students on math performance [73]; therefore, race and gender seemed to act as independent effects in their study.

I. First-generation college students

For this study, we define first-generation college students (FGCS) as students for whom neither parent completed a four-year college degree. First-generation college students enroll in 4-year colleges at a lower rate, attend less prestigious schools correcting for academic preparation, and are retained in college at lower rates than non-first-generation students [74–76]. First-generation college students are also retained within STEM majors at a lower rate than non-FGCS students [77,78]. This may be partially caused by the generally lower socioeconomic status of FGCS students who often cite financial reasons for leaving college [76]. First-generation college students have similar career success to non-FGCS students both in employment and salary; however, they enroll in graduate or professional programs at lower rates than non-FGCS students [74,75].

II. METHODS

A. Instrument

This work reports measurements using the Force and Motion Conceptual Evaluation (FMCE). The FMCE is a 43-item instrument (excluding the energy questions) that measures a student's understanding of Newton's laws and kinematics [20]. Thornton *et al.* introduced a modified scoring of the instrument [79] which produced a total score of 33 by eliminating some items and scoring some blocks of items as all or nothing; we used this scoring method in this work.

B. Sample

This study was performed at a large eastern land-grant university serving approximately 30 000 students. Its overall range of undergraduate ACT scores was 21–26 covering the 25th to the 75th percentile [80]. The university's undergraduate demographics were 80% White, 6% International, 4% African American, 4% Latinx, 4% reporting two or more races, 2% Asian with other groups 1% or less [80].

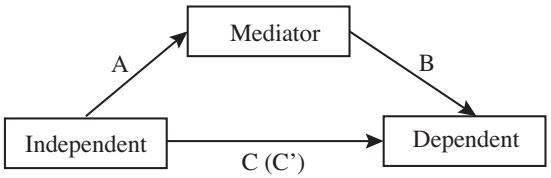

FIG. 1. Mediation process.

The sample was collected in the introductory calculus-based mechanics class serving physical scientists and engineers. Race, ethnicity, gender, and first-generation data were taken from university records. Race and ethnicity were recorded as "American Indian/Alaska Native," "Asian," "Black/African-American," "Hawaiian/Pacific Islander," "Latinx," and "White." Students could select multiple categories. Students were coded as underrepresented ethnic or racial minority status (UERM = 1) if they identified as American Indian/Alaska Native, Black/African-American, Hawaiian/Pacific Islander, or Latinx; otherwise, they were coded as not having underrepresented ethnic or racial minority status (UERM = 0). First-generation college students were coded as FGCS = 1, non-FGCS students as FGCS = 0. Gender was also accessed from university data; the university records gender as a binary variable where the student must elect either male or female. Gender was coded as Gender = 1 if the student identified as female and Gender = 0 if the student identified as male.

Data were collected from the Spring 2011 to the Fall 2019 semester. In that time, 7817 students enrolled in the course. Only students who completed the course for a grade, who had both a FMCE pretest and post-test score, and who had reported ACT or SAT scores to the institution were included in the study producing a sample with $N = 4490$ records.

Course grades were converted to a numeric scale with F = 0 and A = 4. SAT and ACT mathematics scores were converted into percentile scores using scales published by the testing companies. If a student reported both ACT and SAT scores, the percentile scores were averaged. This combination was represented by the variable ACTM because the majority of the students reported ACT scores.

For all regression analyses, all continuous variables were standardized by subtracting the mean and dividing by the standard deviation.

C. Mediation and moderation

Mediation and moderation represent a powerful framework for analyzing variables that influence the relation between an independent variable such as membership in an underrepresented group and a dependent variable such as the grade received in a physics course. We adopt the framework for mediation proposed by Baron and Kenny [81] and summarized by the path model in Fig. 1.

Figure 1 represents the path model for the relations of the dependent variable (Dep), independent variable (Indep), and the mediator (Med); the path model provides a visual summary of a series of linear regressions. The path model is constructed of a series of nodes connected by arrows. The node at the head of the arrow is the dependent variable in the regression. The node at the tail of the arrow represents one independent variable in the regression. The weight of the arrow is the regression coefficient. The full regression equation is recovered by summing all arrows incident on a node, the dependent variable.

The total effect C is measured through the regression

$$\text{Dep} = \beta_1 + C \times \text{Indep} + \epsilon_1, \tag{1}$$

where β_i is the intercept and ϵ_i is the residual error. If the independent variable is a dichotomous demographic variable, the total effect C is the difference between the demographic groups in standard deviation units.

With the mediator, Indep acts through two paths: the Direct Path characterized by C' and the Indirect Path through the mediator composed of a path from Indep to Med (A) and the path from Med to Dep (B). These parameters are measured by

$$\text{Med} = \beta_2 + A \times \text{Indep} + \epsilon_2, \tag{2}$$

$$\text{Dep} = \beta_3 + C' \times \text{Indep} + B \times \text{Med} + \epsilon_3. \tag{3}$$

The mediation is significant if A, B, and C are significant regression coefficients and if $C' < C$. To further demonstrate significant mediation, bootstrapping with 1000 replications was used to show the 95% confidence interval of the total indirect effect ($A \times B$) does not include zero. In this way, the overall effect of the independent variable on the dependent variable, C, is partitioned into the part resulting from the mediator ($A \times B$) and the part not resulting from the mediator (C') such that $C = C' + A \times B$. If the independent variable is a dichotomous demographic variable, C' is the remaining difference between the minority and majority members of the group controlling for the mediating variable.

Moderation occurs when one variable, the moderator, affects the relation of two other variables. To test for moderation, interaction terms are added to the regression equation; an interaction term involves products of independent variables. The moderation of the relationship between the independent variable and the dependent variable by the moderator (Mod) is given by

$$\text{Dep} = \beta_4 + \beta_5 \times \text{Indep} + D \times \text{Indep} \times \text{Mod} + \epsilon_4. \tag{4}$$

The moderation is significant if D is significant. If the independent variable is a dichotomous demographic variable, D represents the difference in the regression coefficient (the slope) between the majority and minority members of the group.

III. RESULTS

Table I presents descriptive statistics for each demographic group. The subgroups of each group are compared with a t test; the effect size of the difference is reported as Cohen's d. The statistical significance of the t test is reported as a superscript on d.

While no significant difference in either course grade or ACTM was measured between men and women, a significant difference in the pretest scores of men and women did exist, a small effect, that grew to near a medium effect size on the post-test. Pretest and post-test differences were smaller between UERM and non-UERM students and between FGCS and non-FGCS students at the small or

TABLE I. Descriptive statistics. All values are the mean ± the standard deviation. Cohen's d measures the effect size of the difference between the majority and minority members of each group. The significance of the difference tested by a t test is presented as a superscript on d. Note superscripts denote (*) $p < 0.1$, (a) $p < 0.05$, (b) $p < 0.01$, and (c) $p < 0.001$.

	N	Physics grade	ACT Math %	Pretest%	Post-test%
Overall	4490	2.95 ± 0.98	80 ± 15	23 ± 18	49 ± 28
			Gender		
Women	995	2.98 ± 0.96	80 ± 15	18 ± 14	38 ± 24
Men	3495	2.95 ± 0.99	80 ± 15	24 ± 19	51 ± 29
d		0.03	0.03	0.33^c	0.47^c
			Underrepresented ethnic or racial minority status		
UERM	335	2.69 ± 1.00	72 ± 18	20 ± 16	42 ± 26
Not UERM	4155	2.98 ± 0.96	80 ± 15	23 ± 18	49 ± 28
d		0.29^c	0.55^c	0.17^c	0.26^c
			First-generation college student status		
FGCS	642	2.80 ± 1.01	76 ± 17	21 ± 16	46 ± 27
Not FGCS	3848	2.98 ± 0.98	80 ± 15	23 ± 19	49 ± 28
d		0.19^c	0.24^c	0.15^c	0.12^b

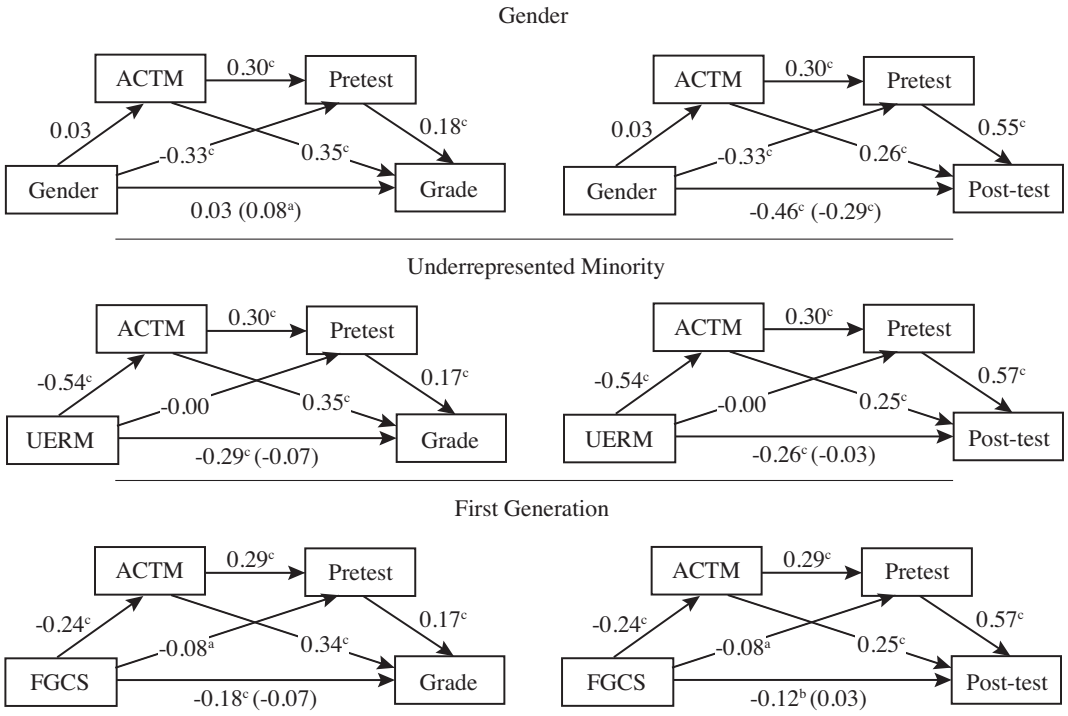

FIG. 2. Path diagrams of the relation of gender, UERM status, and FGCS status, ACTM, pretest scores, post-test scores, and course grades. A path model shows the results of multiple regression analyses. The dependent variable of each regression is the node at the head of the path. The nodes at the tail of the path represent the independent variables. The number on each path is the regression coefficient for the independent variable. For the path with two numbers reported, the number outside of the parenthesis is the regression coefficient using only the demographic variable; the number inside the parenthesis is the regression coefficient with the demographic variable and the other variables included in the model. Note superscripts denote (*) $p < 0.1$, (a) $p < 0.05$, (b) $p < 0.01$, and (c) $p < 0.001$.

less than small effect size level. Significant differences in both course grade and ACTM were measured between UERM and non-UERM students and between FGCS and non-FGCS students; for UERM students the difference in ACTM represented a medium effect.

A. Mediation analysis

The relation between demographic characteristics (Gender, UERM, and FGCS), measures of prior preparation (ACTM and Pretest), and learning outcomes (Post-test and Grade) is shown in Fig. 2. The line between the demographic variable and the outcome variable reports two numbers. The number not in parentheses is the total effect using only the demographic variable and the outcome variable; this is analogous to C in Fig. 1. The total effect represents the difference between the minority and majority members of the group measured in standard deviation units; if the total effect is positive, the minority members perform on average better than the majority members. The result in parentheses is the remaining effect after correcting for ACTM and Pretest; the result after all variables are included in the model and is analogous to C' in Fig. 1. The remaining direct effect C' is the difference between the minority and majority members of the demographic group controlling for both ACTM and pretest score.

The size of the coefficients in Fig. 1 can be understood in terms of effect sizes. For dichotomous independent variables, the coefficient represents the change in the dependent variable between the groups measured in standard deviation units. This quantity is analogous to Cohen's d and may be characterized by its effect size criteria. For a continuous independent variable, the regression coefficient represents the relationship between two normalized continuous variables and is related to the correlation coefficient between those variables. Cohen's criteria for the correlation coefficient is that 0.1 is a small effect, 0.3 is a medium effect, and 0.5 is a large effect [18].

For both UERM and FGCS students, the relation of the demographic variable to grade was similar. Without correcting for ACTM or pretest, the relation between membership in the underrepresented group and course grade was significantly negative (a small effect for UERM students, slightly less than a small effect for FGCS students). This relationship was strongly mediated by pretest scores and ACTM; with these variables in the model, the relationship was no longer significant and was reduced to less than half its initial value (both less than one-half the threshold for a small effect). The relation of gender to grade was dramatically different. The total effect was slightly positive and not significant. The total effect represents the difference between the minority and majority members of the

demographic group; for gender, a positive total effect implies women performed more strongly. Controlling for ACTM and pretest exposed a small, significant positive relationship between gender and grade (less than a small effect).

The relation of the demographic variables and post-test score was fairly similar with a significant negative overall effect of demographics on post-test score for all underrepresented groups (near a medium effect for women, a small effect for UERM students, and less than a small effect for FGCS students). The differences were mediated by ACTM and pretest for all groups; for UERM and FGCS students this mediation reduced the relation between the demographic variable and post-test score substantially until the relation was no longer significant (both much less than a small effect). For women, the relation was reduced somewhat, but remained substantial (-0.29) representing a small effect.

The analysis partitioning the total effect C into a remaining direct effect C' and the indirect effect through the mediator ($A \times B$) can be generalized for the models in Fig. 2. Unlike the simple mediation in Fig. 1, the mediation can now act through any path connecting the demographic variable to the outcome variable. To find the effect along some path, the coefficients of each edge in the path are multiplied together. For example in Fig. 2, the amount of the total effect of UERM status on course grade (-0.29) that acts through the path from UERM to ACTM to Pretest to Grade is $-0.54\times0.30\times0.17=-0.027$, or $100\%\times(-0.027)/(-0.29)=9\%$ of the total effect. Table II summarizes this analysis. For the effect of gender on grade and the effect of FGCS status on post-test, paths have both positive and negative effects. For these graphs, the total effect was calculated as the sum of the absolute values of the paths. No mediation is present in the relation of gender to grade; the percentages for this graph are presented for completeness.

For the course grade of UERM and FGCS students, the paths through pretest account for little of the total effect; the path through ACTM directly to grade accounts for the largest percentage of the total effect. As such, general academic preparation accounts for the largest amount of the difference in course grades for UERM and FGCS students; prior preparation in physics is less important. For FGCS students, more of the total effect remains unexplained by either ACTM or pretest score; for these students a higher percentage of the total effect remains after controlling for ACTM and pretest score.

The relation of gender, UERM status, and FGCS status to post-test score was dramatically different for the three groups. For women, the majority of the total effect (60%) was unexplained by either pretest score or ACTM with most of the remaining effect explained by pretest (38%). For both FGCS and UERM students, ACTM and pretest score explained over 80% of the total effect. For UERM students, the majority of the total effect went through ACTM either directly (52%) or through the effect of ACTM on pretest scores (35%). Very little of the overall effect acted through the pretest alone. As such, for UERM students differences in post-test scores are more related to general high school academic preparation than to specific preparation in physics. For FGCS students, the total effect was fairly evenly distributed through the paths through ACTM and pretest. As such, while pretest score and ACTM are important for explaining post-test and grade differences, their relative importance and the amount of the total effect explained varies widely by subgroup.

B. Within group moderation analysis

The path model in the previous section assumed that the relations between the variables were all linear; that one variable did not moderate the relation between two of the other variables. This assumption can be tested by adding interaction terms to the regression models. The results of this analysis are shown in Figs. 3–5.

The effect of higher order terms in the regression models are represented by the inclusion of variables on the edges of the path models in Figs. 3–5. For example, in Fig. 3 the edge from ACTM to pretest has the value $0.33-0.16\times\text{Gen}$. This indicates that the regression equation with pretest as the dependent variable contains significant terms $0.33\times\text{ACTM}$ and $-0.16\times\text{ACTM}\times\text{Gender}$. The $-0.16\times\text{ACTM}\times\text{Gender}$ term also appears as $-0.16\times\text{ACTM}$ on the edge from gender to pretest. We have included all significant edges without correcting the edges for the number of statistical tests performed. Because

TABLE II. Path analysis of the effect of demographics on course grade and FMCE post-test score. Demo represents one of the dichotomous demographic variables: Gender, UERM, or FGCS. The values represent the percentage of the total effect of the demographic variable on the outcome variable which is explained by each path.

	Outcome = Grade			Outcome = Post − test		
	Gender	UERM	FGCS	Gender	UERM	FGCS
Demo → Outcome	54%	26%	41%	60%	12%	18%
Demo → ACTM → Outcome	6%	64%	45%	2%	52%	33%
Demo → Pretest → Outcome	39%	0%	8%	38%	1%	26%
Demo → ACTM → Pretest → Outcome	1%	9%	7%	1%	35%	22%

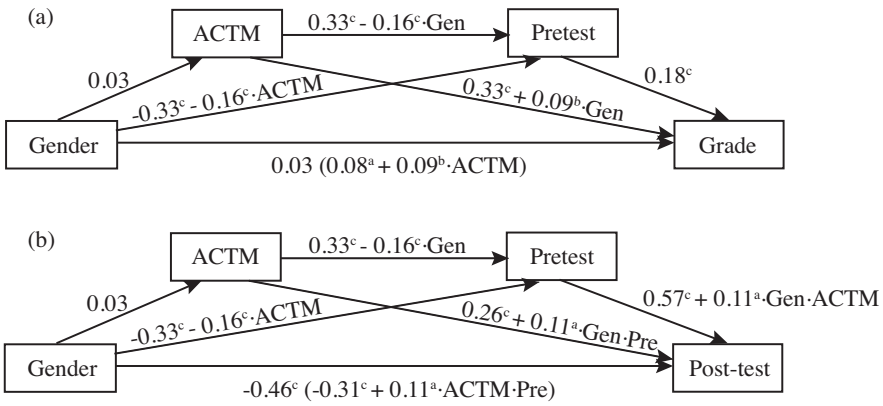

FIG. 3. Moderated mediation examining the effect of gender on grades (a) and post-test score (b). Gender is abbreviated Gen; Pretest as Pre. Note superscripts denote (*) $p < 0.1$, (a) $p < 0.05$, (b) $p < 0.01$, and (c) $p < 0.001$.

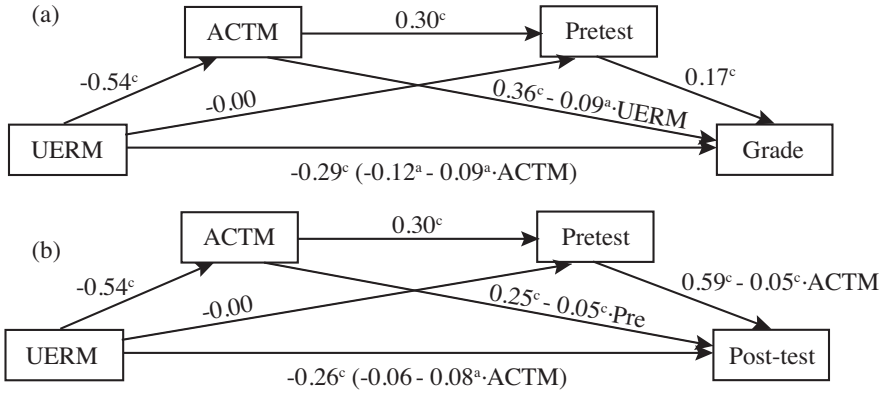

FIG. 4. Moderated mediation examining the effect of underrepresented minority status on grades (a) and post-test score (b). Pretest is abbreviated as Pre. Note superscripts denote (*) $p < 0.1$, (a) $p < 0.05$, (b) $p < 0.01$, and (c) $p < 0.001$.

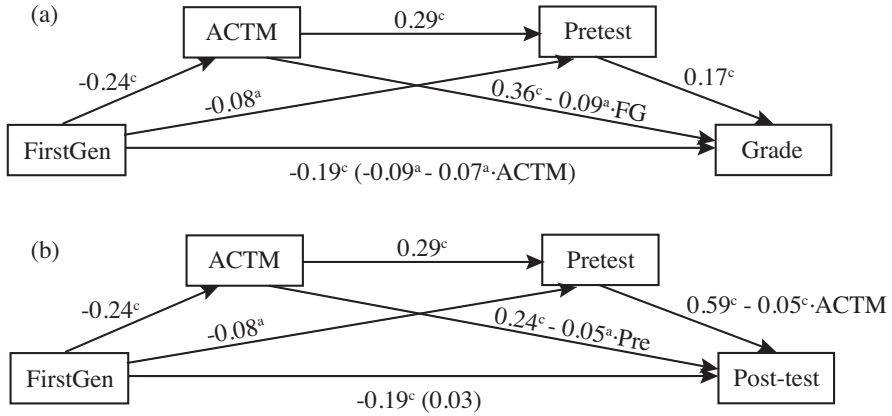

FIG. 5. Moderated mediation examining the effect of first-generation status on grades (a) and post-test score (b). Pretest is abbreviated as Pre. Note superscripts denote (*) $p < 0.1$, (a) $p < 0.05$, (b) $p < 0.01$, and (c) $p < 0.001$.

of the smaller size of the minority demographic groups lowering statistical power, it is more appropriate to examine the effect size of the regression coefficient.

Figures 3–5 show a number of important interactions indicating that many of the relations in Fig. 2 are moderated. Figure 3 shows that gender moderates the relation between ACTM and pretest score, with ACTM less predictive of pretest score for women. The relation of ACTM to grade is also moderated by gender with ACTM more predictive of grade for women. A similar moderation

TABLE III. The number of students in each demographic subgroup.

Gender and UERM		
	Women	Men
UERM	73	262
Not UERM	922	3233
Gender and FGCS		
	Women	Men
FGCS	128	514
Not FGCS	867	2981
UERM and FGCS		
	UERM	Not UERM
FGCS	61	581
Not FGCS	274	3574

of the relation of ACTM to grade is observed for both UERM students and FGCS students. While the interactions vary somewhat between different demographic groups, the main effects for the relation of ACTM to pretest, ACTM to grade, ACTM to post-test, and pretest to post-test are remarkably similar for all groups. The relation of the demographic variables to ACTM, pretest, post-test, and grade continue to be very different.

C. Between group moderation analysis

The moderation analysis of the previous section can be extended to examine multiple demographic groups simultaneously. Mathematically, this is accomplished by introducing product terms of two different dichotomous demographic variables (i.e., UERM × FGCS) in the regression equations. Introducing these product terms has the effect of dividing the sample into the four subgroups defined by all combinations of the levels of the two variables. The combination representing the minority class of both variables often contains many fewer students than the other combinations as shown in Table III where there are only 61 FGCS-UERM students and 73 female UERM students. The low numbers in these subgroups raise the concern of whether the sample has sufficient statistical power to detect these interactions. Exploring these interactions are, however, crucial to understanding the outcomes of students who are members of multiple minority demographic groups.

1. Statistical power

Statistical power represents the likelihood that an analysis can detect an effect if it is actually present. Statistical power may depend on sample size, the significance threshold, and the size of the effect one is attempting to measure. Before exploring the fully interacting models that generalize the models in Fig. 2, the moderation of the total effect C by membership in multiple minority demographic groups was explored. The interacting total effect regression equation for the relation of gender and UERM status to ACTM score is given by

$$\mathrm{ACTM} = \beta_0 + \beta_1 \times \mathrm{Gender} + \beta_2 \times \mathrm{UERM} + D \times \mathrm{Gender} \times \mathrm{UERM} + \epsilon. \quad (5)$$

The moderation is significant if D is significant. A power analysis was conducted to determine what level of D could be reliably detected as significant. Using the significance threshold, $p < 0.05$ and 1000 bootstrap replications, $D = 0.27$ was found significant in 50% of the replications, $D = 0.39$ in 80%, and $D = 0.45$ in 90%. At this threshold, the sample size of female UERM students only allows the identification of interactions nearing Cohen's criteria for a medium effect, $d = 0.5$. Relaxing the threshold of significance to $p < 0.1$ changed these results little; at this threshold, $D = 0.34$ was detected as significant in 80% of the replications. Similar results were obtained for other underrepresented groups.

To allow the exploration of smaller effects, a method was borrowed from the machine learning community. Unbalanced samples are a routine problem in the application of machine learning to classification problems. For example, if one wishes to detect credit card fraud, only a tiny fraction of the overall cases involve fraud [82]. Without correction, the machine learning algorithms are trained overwhelmingly on the majority case and perform poorly on the minority case [83–85]. One method of correcting for sample imbalance is oversampling, randomly creating new minority cases related to the existing cases. This can be done simply by randomly duplicating cases. More sophisticated methods have been developed; we use the synthetic minority oversampling technique (SMOTE) [86] which generates new minority cases by interpolating between cases. SMOTE was used to oversample students in the minority in two of the demographic groups to increase the size of the this subset until it was approximately the size of the next smallest subset (male UERM students when examining gender and UERM status). With oversampling, at the $p < 0.05$ significance threshold, $D = 0.18$ was detected as significant in 50% of the replications, $D = 0.26$ in 80%, and $D = 0.31$ in 90%. At $p < 0.1$, $D = 0.23$ was detected as significant in 80% of the replications. As such, oversampling greatly increased statistical power.

Oversampling by manufacturing additional students by interpolating between already existing students is not optimal and risks overfitting the data. It would be much better to have enough minority cases. Unfortunately, the high degree of underrepresention of both women and UERM students in physics classes would require many years of additional data collection to achieve the power of oversampling. Until these data are collected, it seems better to have some method to study multiply underrepresented students. As such, the results of this section should be considered as suggestive and indicating that additional

TABLE IV. Interaction of demographic variables predicting performance measures without oversampling. Gender is abbreviated Gen. Note superscripts denote (*) $p < 0.1$, (a) $p < 0.05$, (b) $p < 0.01$, and (c) $p < 0.001$.

	Gen × UERM	Gen × FGCS	FGCS × UERM
ACTM	0.32^{a}	−0.04	−0.10
Pretest	0.09	−0.11	0.18
Post-test	0.02	−0.14	0.13
Grade	0.02	−0.16	0.13

research is needed. A possible path forward that avoids the dangers of oversampling but allows analysis in a reasonable time frame is to collect data from multiple universities and perform the same analysis using hierarchical linear modeling which naturally accounts for differences resulting in data collected at multiple institutions.

2. Direct interactions

The moderation of combinations of demographic variables was explored. This analysis seeks to determine if membership in multiple minority demographic groups has additional effects beyond the cumulative effect of differences already explored; for example, is there an additional effect of being a female FGCS student not captured by the effect of being a female student and the effect of being a FGCS student? We first explored the direct effect of each combination of demographic variables on each performance measure: ACTM, Pretest, Post-test, and Grade. The results are shown in Table IV, which reports the D coefficient in Eq. (5) without oversampling. Table V reports the same information with oversampling.

Oversampling changed the value of the coefficients little, but the significance level of the coefficients dramatically. These significance levels are now better aligned with what one might expect by the effect size of the coefficient. In the future, when a power analysis suggests there is insufficient sample size to draw meaningful conclusions about the significance of an effect, the effect sizes should be considered instead. Quantities that are not statistically significant, but are of a substantial effect size, should be discussed as warranting further research.

TABLE V. Interaction of demographic variables predicting performance measures with oversampling. Gender is abbreviated Gen. Note superscripts denote (*) $p < 0.1$, (a) $p < 0.05$, (b) $p < 0.01$, and (c) $p < 0.001$.

	Gen × UERM	Gen × FGCS	FGCS × UERM
ACTM	0.31^{c}	−0.04	−0.08
Pretest	0.06	−0.11	0.16^{*}
Post-test	−0.01	-0.16^{a}	0.09
Grade	−0.00	-0.14^{*}	0.13

The interaction term in Eq. (5) corrects the sum of the effect of being in each demographic group for the effect of being in two underrepresented demographic groups. With the interaction and oversampling, women and men perform equally on the ACTM (a less than $\beta_1 = 0.01$ standard deviation advantage for women), while UERM students have a $\beta_2 = -0.61$ standard deviation disadvantage relative to non-UERM students in ACTM score. There is a $D = 0.31$ standard deviation interaction between gender and UERM status; therefore, UERM women are at less of a disadvantage than UERM men with a total disadvantage of $0.00 - 0.61 + 0.31 = -0.30$ compared to the male UERM disadvantage of $0.00 - 0.61 = -0.61$. First-generation UERM students have a small additional advantage (0.16) over the total disadvantage of either FGCS or UERM students on pretest scores. First-generation women have small additional disadvantage in post-test score (−0.16) and course grade (−0.14) over the total disadvantage of FGCS students and women. These additional advantages or disadvantages may result from different demographic groups suffering advantages or disadvantages from the same sources. For example, it may be that UERM and FGCS students both have less access to enriched high school curricula; the noninteracting regression terms captures the resulting disadvantage in pretest scores for each group but doubly counts the disadvantage for students in both groups. The interaction may be positive to correct for this overcounting.

The interaction of demographic variables with preparation variables was also investigated. These fully interacting models are presented in the Supplemental Material [87] and discussed as part of research question 4.

IV. DISCUSSION

This study sought to answer four research questions. These will be explored in the order proposed.

A. Research questions

RQ1: Are there differences in conceptual post-test scores and course grades between demographic groups in physics classes? The differences between demographic groups were largely what was expected from previous studies. First-generation and UERM students were less well prepared for college STEM courses as measured by ACTM, as has been reported in other studies [3,5,30,31]. Both UERM and FGCS students also had lower course grades, a small effect. Women, FGCS students, and UERM students also had lower FMCE pretest scores; this difference was larger for women. Conceptual differences persisted post-instruction with the difference between men and women growing to near a medium effect size. The pretest and post-test differences for women were largely consistent with the differences reported by Madsen *et al.* [21]. The grade differences for UERM students were consistent with the

differences reported by Hazari *et al.* [6]. The post-test performance differences between majority and nonmajority students were similar to those reported by Brewe *et al.* [24].

RQ2: How are these achievement differences related to general academic preparation and prior preparation in physics? Path analysis was used to examine the relations of demographic variables, course outcomes, and measures of general academic preparation (ACTM) and prior preparation in physics (FMCE pretest score). Differences between UERM students and White non-Latinx students on both course grades and post-test scores were mostly removed by controlling for ACTM; this effect was previously reported by Riegle-Crumb and King [38]. This result is also consistent with Kost, Pollock, and Finkelstein [22] who found ethnicity was not a significant regression coefficient for predicting FMCE post-test score if FMCE pretest score, gender, and mathematics preparation were controlled for. A similar result was found for FGCS students. Salehi *et al.* [17] also reported that differences in final exam scores by gender, UERM status, and FGCS status were mediated by ACTM and pretest scores. This was consistent with the present study for UERM and FGCS students examining grades; however, no difference in physics grade was measured for women in the current study. The strong mediation of the post-test differences of UERM students is also consistent with prior work [24,26] where post-test differences were largely explained by differences in prior preparation.

While ACTM and pretest score explained the differences in course grade and post-test score for FGCS and UERM students, a path analysis showed the two prior preparation factors played a different role for these two groups of students. Table II shows that, for UERM students, the majority of the total effect of underrepresentation on both course grade (73%) and post-test score (87%) acted through ACTM; therefore, general differences in high school preparation were most important for explaining differences in physics outcomes for these students. ACTM could affect the outcome along two paths: directly and through its effect on pretest scores. For UERM students, virtually all the effect of ACTM acted directly on course grade; however, the effect was fairly evenly split between both paths for post-test scores. As such, both general academic preparation and specific preparation in physics were important to explaining post-test differences for UERM students.

The results for FGCS were similar, but more of the total effect of FGCS status on course grade (41%) remained unexplained by either preparation variable. For FGCS students, 52% of the effect on course grade acted through ACTM, 55% for post-test score. Of that acting through ACTM, the vast majority acted directly on course grade with the effect fairly evenly split between the direct path and the path through the pretest for post-test scores. Again, the relative effect of ACTM scores and pretest scores differed by outcome with ACTM most important for course grades and both ACTM and pretest scores important for post-test scores.

This analysis has a number of important implications. Pretest scores measure a combination of physics prior knowledge and general academic preparation, not simply prior physics knowledge. The combination of ACTM and pretest score explains most of the difference in post-test score with only 12% (UERM) and 18% (FCGS) of the total effect unexplained; however, more of the total effect is left unexplained for course grades (UERM 26% and FGCS 41%). For instructors wishing to understand course grades, FGCS students, and to a lesser extent, UERM students are not well characterized by these two variables alone.

While ACTM and pretest score strongly mediated the relation of underrepresentation to course outcomes for UERM and FGCS students, the relations were dramatically different for women. While no difference in course grade or ACTM was measured between men and women, large differences in pretest and post-test scores were measured. The effect of gender varied strongly with the outcome variable. The overall effect of gender on course grade was not significant; controlling for ACTM and pretest revealed a significant ($p < 0.05$, less than a small effect) advantage for women. For post-test score, the total effect (-0.46) was substantial (near a medium effect size). This effect was weakly mediated by ACTM and pretest score reducing the effect to -0.29, still a medium effect. Using path analysis, 60% of the total effect of gender remained unexplained by either ACTM or pretest score. As such, for instructors wishing to understand post-test scores, pretest scores and ACTM leave most of the gender difference unexplained.

The effect of gender on post-test scores was starkly different than that of UERM or FGCS status controlling for academic preparation. In most cases, UERM and FGCS differences were greatly reduced when ACTM and pretest score were added to the models; however, gender differences were only reduced slightly. The failure of prior preparation to mediate gender differences in post-test scores is consistent with a number of other studies reporting similar effects [23,24,26].

The mediation results suggest that the differences in conceptual physics performance by UERM students or FGCS students are largely the result of the prior academic preparation of the student. The failure to observe the same mediation of the regression coefficients for women suggests that the source of performance differences by gender are different than the sources of performance differences by UERM or FGCS status. Many possible sources of performance differences by gender other than academic preparation have been explored including instrumental bias [64], stereotype threat [56–58], math [51,52] and science anxiety [53–55], and mode of instruction [24,65–67].

Salehi *et al.* [17] reported a similar path model for the effect of gender on final exam score, but performed the mediation analysis using structural equation models which

do not provide the partitioning of the overall effect shown in Table II. For this comparison, only the largest sample in Salehi *et al.*, labeled PM in that work, is considered. The effect of gender on course grade was very different than that reported in the current work. While women had slight advantages in course grade and ACTM scores in the current work, women in Salehi *et al.* had significantly lower final exam and ACTM scores. No mediation was detectable for women in the present study because the total effect of gender on course grade was not significant; in Salehi *et al.*, ACTM slightly mediated the effect of gender on final exam scores while FMCE pretest score strongly mediated the effect.

As such, the relation of membership in a minority demographic group to course outcomes was inconsistently related to prior preparation measures.

RQ3: Are these relations consistent for the minority and majority members of the demographic groups? The relations between preparation measures and outcomes differed for some minority members of each demographic group; the relations were moderated by membership in the group. Some preparation measures moderated the effects of other measures. These relationships are summarized in Figs. 3–5. This analysis revealed a number of important relations. ACTM scores were less predictive of pretest score but more predictive of course grade for women; while ACTM scores were less predictive of course grade for UERM students and FGCS students. Relations between ACTM, pretest, and course outcomes were often moderated with ACTM scores modifying the relation of pretest to post-test. This moderation caused pretest scores to be less predictive of post-test scores for UERM and FGCS students with higher ACTM scores. The moderation was different for men and women; pretest score was more predictive of post-test scores, but only for women with higher ACTM scores.

Some of these results are directly comparable to those of Salehi *et al.* [17] which uses final exam score as a dependent variable. For this comparison, only the largest sample in Salehi *et al.*, labeled PM in that work, is considered. As in this work, there was a significant Gender × ACTM interaction with coefficient 0.104 in the saturated model (Model 14 in Salehi *et al.*, Appendix Table VI). This is very similar to the Gender × ACTM interaction shown in Fig. 3 where the interaction has coefficient 0.09 with dependent variable course grade. Salehi *et al.* only considered interactions on edges in the path model ending on the dependent variable and, therefore, other interactions found in this work could not be compared.

Salehi *et al.* [17] also considered interacting models using all three demographic variables used in this work: gender, UERM status, and FGCS status. While the current work considered interactions between these variables in pairs, some results are comparable. Salehi *et al.* identified a strong FGCS × ACTM interaction (coefficient −0.115, $p < 0.001$) and a weaker ACTM × Pretest interaction (coefficient 0.046, $p < 0.05$). The current work identified a similar FGCS × ACTM interaction (coefficient −0.09, $p < 0.05$) predicting course grade, but did not detect the ACTM × Pretest interaction. The current study also identified a ACTM × Pretest interaction for both UERM and FGCS students predicting post-test score.

RQ4: Are there additional effects of being a member of multiple underrepresented groups? Tables IV and V present the interaction of various combinations of demographic groups on the relation of membership in the group and academic preparation and outcome variables. Without oversampling, only one interaction was significant and represented a small effect, the interaction between UERM status and gender. This interaction indicates that UERM women do not experience the sum of the advantage of women and the disadvantage of UERM students on ACTM score. Three other interactions, which only became significant with oversampling, approach the threshold of a small effect. Being both a FGCS student and an UERM student had a positive effect on pretest score while being a FGCS woman had a negative effect on both post-test score and grade. The failure to find a significant gender-UERM interaction for course grade and post-test score was consistent with previous research showing that the effect of these variables is generally independent [73].

The fully interacting models shown in Figs. 1–6 in the Supplemental Material [87] paint a complex picture of the interconnections of being multiply in the minority, academic preparation, and academic outcomes. These were discussed in detail in the Supplemental Material [87]. Beyond the direct effects discussed in the previous paragraph, a number of additional important interactions were identified. The relation of pretest to course grade was different for female FGCS students with pretest more predictive of grade, an effect that also depended on ACTM scores. The relation of pretest to post-test was also different for these students. ACTM scores were less predictive of grade for UERM women, this effect also depended on pretest scores. For female UERM students, ACTM scores were more predictive of pretest scores. In general, these relations were complex suggesting general conclusions drawn from an aggregated sample may apply with differing accuracy to demographic subgroups and to members of those groups with different levels of academic preparation.

B. Additional observations

This work examined the interaction of multiple identities held by students who are substantially underrepresented in physics and introductory physics classrooms: gender, UERM status, and FGCS status. Intersectionality theory holds that these and other social identities intersect and interact to produce patterns of discrimination (or privilege). Three broad cases were possible for the interaction of these identities. The first case is that there is no interaction between membership in different minority demographic

groups; the variables act independently. In general, this was the case as shown in Table IV where only one interaction represented a small effect size. A multiply underrepresented student accumulates disadvantages (or a multiply overrepresented student advantages) that are the sum of individual effects of UERM status, FGCS status, and gender. In the second case, multiple underrepresentation could be compensatory where being multiply underrepresented did not produce the sum of the effects of the individual underrepresentations. This was true for the ACTM scores of UERM women (a small effect) and first-generation UERM students for pretest scores (near a small effect). This is the most positive case. The third case is a multiplicative effect where being multiply underrepresented produces an additional effect beyond the sum of the individual effects of underrepresentation. This study found evidence of a multiplicative effect for FGCS women on both post-test scores and course grades (near a small effect).

We acknowledge that intersectionality theory is rooted in Black feminism and shares tenants with critical theories which value lived experience over quantitative generalizations [88]. Thus, a quantitative study may not be suited for understanding the complexities of how multiple identities interact and relate to student performance. Nevertheless, the finding that the effects of underrepresentation are generally additive implies multiply underrepresented students are particularly at risk.

This study utilized an *equity of parity* model [8] which defines equity as closing differences in students' prior preparation on postinstruction performance measures such as post-test scores. This model is useful in that it measures the ability to support students of diverse groups despite prior disadvantages based on identity. The shortcoming of this model is that it relies on students from disadvantaged or marginalized groups having larger gains than those from comparison groups. However, as suggested by Rodriguez *et al.* [8], the key to successfully utilizing this model may be to focus on opportunity and access and eliminate differences before they appear. This study contributes to that goal by identifying prior preparation as a predictor for the differences seen in performance on conceptual tests and course grades. Future studies should use these markers to design interventions to prevent differences in post-test scores and course grades from arising and to identify and address factors that create differences in prior preparation.

This study presented multiple comparisons with Salehi *et al.* [17]. While some features were consistent, such as the mediation of FGCS and UERM students' final exam score or course grade; many were not. This suggests that many of the details of the relation of underrepresentation to physics outcomes depend on specific features of the student population.

V. IMPLICATIONS

The results of this work have implications for three broad communities: physics instructors in general, physics instructors seeking to use academic preparation variables for instructional decisions, and PER researchers. For instructors in general, they paint a complex picture of how differences in preparation of students affect outcomes. These differences are present in every physics class. That general academic preparation (ACTM) and specific preparation in physics (FMCE pretest score) are differentially important for some demographic groups suggest that interventions and other instructional decisions designed to help students must address multiple dimensions of under preparation. For instructors seeking to use academic preparation variables or pretest scores for instructional decisions, the observation that the relation of these variables to course outcomes are moderated suggest some measures will be less accurate predictors for some subgroups and that general models constructed based on an aggregated dataset dominated by majority students will be more accurate for those students than for some students in minority demographic groups. For PER researchers, the observation that 26% (UERM), 41% (FGCS), and 54% (gender) of the difference in course grade and 12% (UERM), 18% (FCGS), and 60% (gender) of post-test score differences remains unexplained by either pretest or ACTM suggest additional or refined instruments are required to characterize the initial state of preparation of physics students and the instructional setting.

VI. RECOMMENDATIONS

This study detected differences in both general academic preparation and specific preparation in physics for multiple demographic groups in the minority in most physics classes. Most universities maintain substantial student demographic and academic data. This data should be used to develop a general overview of the academic characteristics of the students in introductory physics classes to determine the extent to which strong differences in preparation measured by traditional academic markers such as ACTM scores or pretest scores exist. If preparation differences are identified, the department or instructor should prioritize interventions aiding all underprepared students and strive to create classroom structures, assessment, and culture that values the variety of resources that students bring to the classroom. These interventions may include assignments that provide additional practice or support to students who do not achieve a certain level of mastery: adaptive systems, rather than one-size-fits-all assignments. Beyond helping all students to succeed, the current study shows that these interventions are necessary so that differences in preparation do not generate differences in outcomes for women, underrepresented ethnic or racial minority, and first-generation college

students. By implementing instruction that allows all students to succeed, physics classes can provide an equitable learning environment for all students. If inequities persist through physics classes, this may hinder efforts to make the physics profession more inclusive.

VII. LIMITATIONS

This study extended the work of Salehi *et al.* [17] by examining a student population less well prepared for college measured by ACTM and pretest scores. As Salehi *et al.* implored, this work should be extended to many other institutions to understand to what extent the results generalize. This study used an oversampling method when examining interactions between multiple demographic variables. This is not optimal and those results must be viewed as suggestive. Even larger studies are needed to confirm the oversampling results.

VIII. CONCLUSIONS

This study examined the mediation and moderation of the relation of membership in an underrepresented group and course grade or FMCE post-test score by general academic preparation measured by ACT or SAT mathematics percentile scores and prior preparation in physics measured by FMCE pretest score. In general, these variables significantly mediated the relationship; however, the relationship between course grade and gender was not mediated because the total effect was not significant. The percentage of the total effect which could be explained by various paths through the models varied strongly between groups. For UERM and first-generation college students, the majority of the effect of underrepresentation on course grade was explained by the path through general academic preparation (ACTM) while little of the effect was explained by pretest scores. For UERM students this was also true of the effect on post-test scores; however, for FGCS students more of the effect of underrepresentation acted through FMCE pretest scores. The results for gender and post-test score were dramatically different with the majority of the effect of underrepresentation unexplained by either ACTM or pretest score. Significant moderation was also detected indicating that the relation of underrepresentation, ACTM, pretest, post-test, and course grade is not the same for all demographic groups. As such, inferences made about students from ACT or SAT scores and physics pretest scores and actions taken on those inferences, such as lab group placements or the need for educational interventions, may not be equally accurate for students in some demographic groups.

ACKNOWLEDGMENTS

This work was supported the National Science Foundation Grants No. ECR-1561517 and No. HRD-1834569. Some data collection for this project were supported in part by the National Science Foundation under Grant No. EPS-1003907.

[1] B. C. Cunningham, K. M. Hoyer, and D. Sparks, *Gender Differences in Science, Technology, Engineering, and Mathematics (STEM) Interest, Credits Earned, and NAEP Performance in the 12th Grade (NCES 2015-075)* (U.S. Department of Education, National Center for Education Statistics, Institute of Education Sciences, Washington, DC, 2015).

[2] C. Nord, S. Roey, R. Perkins, M. Lyons, N. Lemanski, J. Brown, and J. Schuknecht, *The Nation's Report Card: America's High School Graduates (NCES 2011–462)* (U.S. Department of Education, National Center for Education Statistics, Institute of Education Sciences, Washington, DC, 2011).

[3] P. R. Aschbacher, E. Li, and E. J. Roth, Is science me? High school students' identities, participation and aspirations in science, engineering, and medicine, J. Res. Sci. Teach. **47**, 564 (2010).

[4] *The ACT Profile Report—National Graduating Class 2016* (ACT Inc., Iowa City, IA, 2016).

[5] National Science Board, *Science and Engineering Indicators 2016 (NSB 16–01)* (National Science Foundation, Arlington, VA, 2016).

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

[7] A. N. Parks and M. Schmeichel, Obstacles to addressing race and ethnicity in the mathematics education literature, J. Res. Math. Educ. **43**, 238 (2012).

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

[9] N. Lykke, *Feminist Studies: A Guide to Intersectional Theory, Methodology and Writing* (Routledge, New York, NY, 2010).

[10] M. C. Parent, C. DeBlaere, and B. Moradi, Approaches to research on intersectionality: Perspectives on gender, LGBT, and racial/ethnic identities, Sex Roles **68**, 639 (2013).

[11] K. Rosa and F. M. Mensah, Educational pathways of Black women physicists: Stories of experiencing and overcoming obstacles in life, Phys. Rev. Phys. Educ. Res. **12**, 020113 (2016).

[12] Z. Hazari, P. M. Sadler, and G. Sonnert, The science identity of college students: Exploring the intersection of gender, race, and ethnicity, J. Coll. Sci. Teach. **42**, 82 (2013), https://www.jstor.org/stable/43631586?seq=1.

[13] M. Ong, Body projects of young women of color in physics: Intersections of gender, race, and science, Soc. Probl. **52**, 593 (2005).

[14] L. T. Ko, R. R. Kachchaf, A. K. Hodari, and M. Ong, Agency of women of color in physics and astronomy: Strategies for persistence and success, J. Women Minorities Sci. Eng. **20**, 171 (2014).

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

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

[17] S. Salehi, E. Burkholder, G. P. Lepage, S. Pollock, and C. Wieman, Demographic gaps or preparation gaps?: The large impact of incoming preparation on performance of students in introductory physics, Phys. Rev. Phys. Educ. Res. **15**, 020114 (2019).

[18] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences* (Academic Press, New York, NY, 1977).

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

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

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

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

[23] 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, Phys. Rev. Phys. Educ. Res. **15**, 010131 (2019).

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

[25] R. Henderson and J. Stewart, Racial and ethnic bias in the Force Concept Inventory, in *Proceedings of the 2017 Physics Education Research Conference, Cincinnati, OH* (AIP, New York, 2017), pp. 172–175.

[26] R. Henderson, C. Zabriskie, and J. Stewart, Rural and first generation performance differences on the Force and Motion Conceptual Evaluation, in *Proceedings of the 2018 Physics Education Research Conference, Washington, DC* (AIP, New York, 2018).

[27] U. S. Census Bureau, Washington, DC, Race, https://www.census.gov/topics/population/race/about.html. Accessed 6/13/2018.

[28] A. W. Radford, E. D. Velez, A. Bentz, T. Lew, and N. Ifill, *First-Time Postsecondary Students in 2011-12: A Profile* (U.S. Department of Education, National Center for Education Statistics, Institute of Education Sciences, Washington, DC, 2016).

[29] 2015-16 National Postsecondary Student Aid Study (NPSAS:16) (U.S. Department of Education, National Center for Education Statistics, Institute of Education Sciences, Washington, DC, 2018).

[30] J. M. Harackiewicz, C. S. Rozek, C. S. Hulleman, and J. S. Hyde, Helping parents to motivate adolescents in mathematics and science: An experimental test of a utility-value intervention, Psychol. Sci. **23**, 899 (2012).

[31] L. Musu-Gillette, J. Robinson, J. McFarland, A. KewalRamani, A. Zhang, and S. Wilkinson-Flicker, *Status and Trends in the Education of Racial and Ethnic Groups 2016 (NCES 2016–007)* (U.S. Department of Education, National Center for Education Statistics, Institute of Education Sciences, Washington, DC, 2016).

[32] President's Council of Advisors on Science and Technology, *Report to the President. Engage to Excel: Producing One Million Additional College Graduates with Degrees in Science, Technology, Engineering, and Mathematics* (Executive Office of the President, Washington, DC, 2012).

[33] S. B. Robbins, K. Lauver, H. Le, D. Davis, R. Langley, and A. Carlstrom, Do psychosocial and study skill factors predict college outcomes? A meta-analysis, Psychol. Bull. **130**, 261 (2004).

[34] B. Toven-Lindsey, M. Levis-Fitzgerald, P. H. Barber, and T. Hasson, Increasing persistence in undergraduate science majors: A model for institutional support of underrepresented students, CBE Life Sci. Educ. **14**, 1 (2015).

[35] G. Crisp, A. Nora, and A. Taggart, Student characteristics, pre-college, college, and environmental factors as predictors of majoring in and earning a STEM degree: An analysis of students attending a Hispanic serving institution, Am. Educ. Res. J. **46**, 924 (2009).

[36] R. Koenig, Minority retention rates in science are sore spot for most universities, Science **324**, 1386 (2009).

[37] S. Hurtado, K. Eagan, and M. Chang, *Degrees of Success: Bachelor's Degree Completion Rates among Initial STEM Majors* (Higher Education Research Institute at UCLA, Cooperative Institutional Research Program, Los Angeles, CA, 2010).

[38] C. Riegle-Crumb and B. King, Questioning a White male advantage in STEM: Examining disparities in college major by gender and race/ethnicity, Educ. Res. **39**, 656 (2010).

[39] C. Riegle-Crumb, B. King, E. Grodsky, and C. Muller, The more things change, the more they stay the same? Prior achievement fails to explain gender inequality in entry into STEM college majors over time, Am. Educ. Res. J. **49**, 1048 (2012).

[40] J. Trusty, K. Ng, and M. Plata, Interaction effects of gender, Career Dev. Q. **49**, 45 (2000).

[41] J. L. Kobrin, V. Sathy, and E. J. Shaw, *A Historical View of Subgroup Performance Differences on the SAT Reasoning Test* (The College Board, New York, NY, 2007).

[42] *A Snapshot of the Individuals Who Took the GRE General Test* (Educational Testing Service, Princeton, NJ, 2016).

[43] D. Voyer and S. D. Voyer, Gender differences in scholastic achievement: A meta-analysis, Psychol. Bull. **140**, 1174 (2014).
[44] B. C. Cunningham, K. M. Hoyer, and D. Sparks, *The Condition of STEM 2016* (ACT Inc., Iowa City, IA, 2016).
[45] P. M. Sadler and R. H. Tai, Success in introductory college physics: The role of high school preparation, Sci. Educ. **85**, 111 (2001).
[46] D. F. Halpern, *Sex Differences in Cognitive Abilities*, 4th ed. (Psychology Press, Francis & Tayler Group, New York, NY, 2012).
[47] R. A. Lippa, M. L. Collaer, and M. Peters, Sex differences in mental rotation and line angle judgments are positively associated with gender equality and economic development across 53 nations, Archives of sexual behavior **39**, 990 (2010).
[48] Y. Maeda and S. Y. Yoon, A meta-analysis on gender differences in mental rotation ability measured by the Purdue Spatial Visualization Tests: Visualization of Rotations (PSVT: R), Educ. Psychol. Rev. **25**, 69 (2013).
[49] J. S. Hyde and M. C. Linn, Gender differences in verbal ability: A meta-analysis., Psychol. Bull. **104**, 53 (1988).
[50] E. A. Maylor, S. Reimers, J. Choi, M. L. Collaer, M. Peters, and I. Silverman, Gender and sexual orientation differences in cognition across adulthood: Age is kinder to women than to men regardless of sexual orientation, Archives of sexual behavior **36**, 235 (2007).
[51] X. Ma, A meta-analysis of the relationship between anxiety toward mathematics and achievement in mathematics, J. Res. Math. Educ. **30**, 520 (1999).
[52] N. M. Else-Quest, J. S. Hyde, and M. C. Linn, Cross-national patterns of gender differences in mathematics: A meta-analysis, Psychol. Bull. **136**, 103 (2010).
[53] J. V. Mallow, A science anxiety program, Am. J. Phys. **46**, 862 (1978).
[54] J. V. Mallow and S. L. Greenburg, Science anxiety: Causes and remedies, J. Coll. Sci. Teach. **11**, 356 (1982).
[55] J. Mallow, H. Kastrup, F. B. Bryant, N. Hislop, R. Shefner, and M. Udo, Science anxiety, science attitudes, and gender: Interviews from a binational study, J. Sci. Educ. Technol. **19**, 356 (2010).
[56] H. D. Nguyen and A. Ryan, Does stereotype threat affect test performance of minorities and women? A meta-analysis of experimental evidence, J. Appl. Psych. **93**, 1314 (2008).
[57] G. Stoet and D. C. Geary, Can stereotype threat explain the gender gap in mathematics performance and achievement? Rev. Gen. Psychol. **16**, 93 (2012).
[58] G. M. Walton and S. J. Spencer, Latent ability grades and test scores systematically underestimate the intellectual ability of negatively stereotyped students, Psychol. Sci. **20**, 1132 (2009).
[59] J. G. Cromley, T. Perez, T. W. Wills, J. C. Tanaka, E. M. Horvat, and E. T. Agbenyega, Changes in race and sex stereotype threat among diverse STEM students: Relation to grades and retention in the majors, Contemp. Educ. Psychol. **38**, 247 (2013).
[60] L. McCullough and D. E. Meltzer, Differences in male/female response patterns on alternative-format versions of FCI items, in *Proceedings of the 2001 Physics Education Research Conference*, edited by K. Cummings, S. Franklin, and J. Marx (AIP, New York, 2001), pp. 103–106.
[61] L. McCullough, Gender, context, and physics assessment, J. Int. Womens Studies **5**, 20 (2004).
[62] R. D. Dietz, R. H. Pearson, M. R. Semak, and C. W. Willis, Gender bias in the Force Concept Inventory?, AIP Conf. Proc. **1413**, 171 (2012).
[63] S. Osborne Popp, D. Meltzer, and M. C. Megowan-Romanowicz, Is the Force Concept Inventory biased? Investigating differential item functioning on a test of conceptual learning in physics, in *Proceedings of the 2011 American Educational Research Association Conference* (American Education Research Association, Washington, DC, 2011).
[64] 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).
[65] R. J. Beichner and J. M. Saul, Introduction to the SCALE-UP (Student-Centered Activities for Large Enrollment Undergraduate Programs) project, in *Invention and Impact: Building Excellence in Undergraduate Science*, Technology, Engineering and Mathematics (STEM) Education (American Association for the Advancement of Science, Washington, DC, 2003), pp. 61–66.
[66] M. Lorenzo, C. H. Crouch, and E. Mazur, Reducing the gender gap in the physics classroom, Am. J. Phys. **74**, 118 (2006).
[67] E. Mazur, *Peer Instruction: A User's Manual* (Prentice Hall, Upper Saddle River, NJ, 1997).
[68] S. J. Pollock, N. D. Finkelstein, and L. E. Kost, Reducing the gender gap in the physics classroom: How sufficient is interactive engagement?, Phys. Rev. Phys. Educ. Res. **3**, 010107 (2007).
[69] M. J. Cahill, K. M. Hynes, R. Trousil, L. A. Brooks, M. A. McDaniel, M. Repice, J. Zhao, and R. F. Frey, Multiyear, multi-instructor evaluation of a large-class interactive-engagement curriculum, Phys. Rev. Phys. Educ. Res. **10**, 020101 (2014).
[70] N. I. Karim, A. Maries, and C. Singh, Do evidence-based active-engagement courses reduce the gender gap in introductory physics?, Eur. J. Phys. **39**, 025701 (2018).
[71] R. Henderson, G. Stewart, J. Stewart, L. Michaluk, and A. Traxler, Exploring the gender gap in the Conceptual Survey of Electricity and Magnetism, Phys. Rev. Phys. Educ. Res. **13**, 020114 (2017).
[72] A. L. Traxler, X. C. Cid, J. Blue, and R. Barthelemy, Enriching gender in physics education research: A binary past and a complex future, Phys. Rev. Phys. Educ. Res. **12**, 020114 (2016).
[73] T. Scafidi and K. Bui, Gender similarities in math performance from middle school through high school, J. Instr. Psychol. **37**, 252 (2010).
[74] E. T. Pascarella, C. T. Pierson, G. C. Wolniak, and P. T. Terenzini, First-generation college students: Additional evidence on college experiences and outcomes, J. High. Educ. **75**, 249 (2004).
[75] E. F. Cataldi, C. T. Bennett, and X. Chen, *First-generation students: College access, persistence, and postbachelor's outcomes* (National Center For Education Statistics, Washington, DC, 2018).

[76] J. Redford and K. M. Hoyer, *First-generation and continuing-generation college students: A comparison of high school and postsecondary experiences* (National Center For Education Statistics, Washington, DC, 2018).

[77] X. Chen, *STEM attrition: College students' paths into and out of STEM fields* (National Center For Education Statistics, Washington, DC, 2013).

[78] D. Verdin and A. Godwin, First in the family: A comparison of first-generation and non-first-generation engineering college students, in *Proceedings of the Frontiers in Education Conference (FIE), 2015 IEEE* (IEEE, Bellingham, WA, 2015), pp. 1–8.

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

[80] US News & World Report: Education, US News, and World Report, Washington, DC, https://premium.usnews.com/best-colleges. Accessed 4/30/2017.

[81] R. M. Baron and D. A. Kenny, The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations, J. Personality Social Psychol. **51**, 1173 (1986).

[82] A. Greeen, *Hands-On Machine Learning with Scikit-Learn & TensorFlow* (O'Reilly, Boston, MA, 2017).

[83] H. He and E. A. Garcia, Learning from imbalanced data, IEEE Trans. Knowl. Data Eng. **21**, 1263 (2009).

[84] N. V. Chawla, N. Japkowicz, and A. Kotcz, Special issue on learning from imbalanced data sets, ACM Sigkdd Explor. Newsl. **6**, 1 (2004).

[85] N. V. Chawla, *Data mining for imbalanced datasets: An overview, in Data Mining and Knowledge Discovery Handbook* (Springer, Boston, MA, 2009), pp. 875.

[86] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, SMOTE: Synthetic minority over-sampling technique, J. Artif. Intell. Res. **16**, 321 (2002).

[87] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevPhysEducRes.17.010107 [URL] for the fully interacting moderated-mediation analysis.

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