Differences in the predictive power of pretest scores of students underrepresented in physics

Dona Sachini Hewagallage and John Stewart
Department of Physics and Astronomy, West Virginia University, Morgantown, WV, USA

Rachel Henderson
Department of Physics and Astronomy, Michigan State University, East Lansing, MI, USA

This study examines the correlation of physics conceptual inventory pretest scores with post-instruction achievement measures (post-test scores, test averages, and course grades). The correlation for demographic groups in the minority in the physics classes studied (women, underrepresented racial/enthic students, first generation college students, and rural students) were compared with their majority peers. Three conceptual inventories were examined: the Force and Motion Conceptual Evaluation (FMCE) ($N = 2450$), the Force Concept Inventory (FCI) ($N = 2373$) and the CSEM ($N_1 = 1796$, $N_2 = 2537$). While many of the correlations were similar, for some of the demographic groups, the correlations were substantially different. There was little consistency in the differences measured. In most cases where the correlations differed, the correlation for the group in the minority was the smaller. As such, pretest scores may not predict course performance for some minority demographic groups as accurately as they predict outcomes for majority students. The pattern of correlation differences did not appear to be related to the size of the pretest score. If pretest scores are used for instructional decisions that have academic consequences, instructors should be aware of these potential inaccuracies and ensure the pretest used is equally valid for all students.

I. INTRODUCTION

Physics instructors and physics education researchers often use the practice of applying a conceptual physics inventory as a pretest and post-test to assess student learning in physics classes [1]. Widely used research-based conceptual physics inventories include the Force Concept Inventory (FCI) [2], the Force and Motion Conceptual Evaluation (FMCE) [3], and the Conceptual Survey of Electricity and Magnetism (CSEM) [4]. Beyond assessing learning, pretest scores are sometimes used for other purposes such as establishing laboratory or recitation groups [2]. As such, pretest scores can directly affect student performance in some physics classes. Beyond their potential effect on learning, pretest scores are also used as independent variables in Physics Education Research (PER) studies to control for prior preparation [5].

A substantial literature suggests students underrepresented in physics do not perform equally on physics conceptual pretests or post-tests. Many studies have found men outperform women on both conceptual mechanics and conceptual electricity and magnetism evaluations [6]. Many explanations have been advanced to explain these differences including prior preparation [5, 7], sociocultural factors [8], and instrumental fairness [9, 10].

Substantially less research has examined the performance of racially or ethnically underrepresented students on physics conceptual inventories. Hazari, Tai, and Sadler reported differences in physics grades by race and ethnicity [11]. Racial and ethnic differences in conceptual inventory performance have also been reported [5, 12–15]. In general, post-test differences have been mediated by prior academic preparation measured by either pretest score or ACT/SAT score. These performance differences are generally consistent with differences in more general measures of academic achievement such as standardized test scores, college enrollment rates and graduation rates [16, 17]. Both women and underrepresented ethnic/racial minority students leave Science, Technology, Engineering, Mathematics (STEM) majors at a higher rate than other students [18–20].

Little work has examined other demographic groups. Henderson, Zabriskie, and Stewart [21] reported that first generation college students (FGCS) and rural students also demonstrated conceptual performance differences. The differences between FGCS and non-FGCS were mediated by ACT/SAT scores; the differences between rural and non-rural students were not [21]. First generation college students are less likely to enroll in college and persist in college at lower rates than non-FGCS students [22–24]. FGCS students are also less likely to be retained in STEM majors [25, 26].

Achievement differences between rural and urban students are less consistent. Few differences in mathematics achievement are reported between rural and urban students [27]; however, rural students are reported to have less access to college preparatory high school classes [28]. Rural students also have lower persistence rates in college than non-rural students [29].

Because pretest scores sometimes are used to make decisions such as the assignment to lab groups that directly affect course grades and are often used in research that has broader academic consequences, it is important that they are equally accurate for all groups of students. Recent research has identified items within many popular conceptual inventories that are unfair to either men or women [9, 10] with the FCI containing more unfair items than the FMCE or the CSEM. A similar fairness analysis has not been performed for other groups and, therefore, the fairness of these instruments for other groups underrepresented in physics classes is unknown. Recent research has also challenged the accuracy of pretest scores as a measure of the prior preparation of women. Henderson *et al.* showed that CSEM pretest score was more weakly correlated with a variable measuring conceptual knowledge not explained by quantitative knowledge for women than for men [30]. They proposed that this resulted from the slightly lower pretest score for women shifting the distribution of female pretest scores closer to the pure guessing distribution. As such, it was more difficult to statistically distinguish moderately prepared women from moderately prepared men. Henderson, Stewart, and Traxler showed that the part of the gender gap attributable to prior preparation was substantially changed if measures of conceptual understanding beyond pretest scores were included [31]. Again, no similar research has been performed with other groups.

This study will address the following research questions: *RQ1: Are pretest scores equally correlated with measures of physics achievement for different demographic groups? RQ2: Are any differences identified related to the magnitude of the pretest score?*

RQ2 explores the conjecture of Henderson *et al.* [30] that if a subgroup of a physics class has pretest scores lower than the majority group, then the group's pretest score may not be as strongly related to other academic measures.

II. METHODS

Sample: Data were collected from the introductory calculus-based physics classes at two large land-grant universities in the US. University 1 had an overall undergraduate population that was 79% White, 7% international, 4% African American, 4% Hispanic, and 4% two or more races, with other groups 1% or fewer [32]. Data were collected from spring 2011 to spring 2019 at University 1. University 2 had an overall undergraduate population that was 76% White, 9% Hispanic, 4% African American, and 4% two or more races, with other groups each 3% or fewer [32]. Data were collected from spring 2003 to fall 2012 semester. For University 1, the FMCE was administered in the introductory mechanics class; the results form Sample FMCE-1 ($N = 2450$) where the number represents in institution. The CSEM was administered in the introductory electricity and magnetism classes forming Sample CSEM-1 ($N = 1796$). For University 2, the FCI was administered in introductory mechanics

to form Sample FCI-2 ($N = 2373$) and the CSEM in introductory electricity and magnetism to form Sample CSEM-2 ($N = 2537$). Rural/non-rural data were not available for all students at University 1.

The samples are related, but not identical, to samples published in earlier works (FMCE-1 [10, 13, 21]; CSEM-1 [10, 13]; FCI-2 [9, 13]; CSEM-2 [10, 13, 30]); additional details of sample collection, institutional setting, and instructional environment can be found in these works. For University 1, the samples differ from prior analysis because they include an additional year of data, but are restricted to matched pretest/post-test pairs. For University 2, the sample differs from previous work by the restriction to matched pretest/post-test pairs.

Measures: Four measures of academic achievement were collected. Conceptual inventory pretest scores, post-test scores, and in-semester test averages were converted to percentages. Course letter grades were converted to a numeric scale with "A" = 4 and "F" = 0.

For University 1, FGCS status, gender, and race/ethnicity data were collected from university records. Gender was coded dichotomously as male and female. Race/ethnicity was dichotomously coded with White non-Hispanic students coded as White and all other students coded as non-White. This dichotomous coding of race/ethnicity and gender is not optimal and obscures the complex and nuanced nature of both constructs [33]; this coding was consistent with that used by the institution for gender and was necessitated by the low number of racial or ethnic minority students. Future studies should be performed to provide a more nuanced treatment of these constructs.

To establish rural status, the student's high school code, the College Entrance Examination Board (CEEB) code, was accessed from university records. This was converted to a National Center for Education Statistics (NCES) school identifier [34]. This allowed the NCES classification of the urbanicity or rurality of the community in which the school is located, the locale code, to be determined for each student [35]. The locale code is measured on a 12-point scale; we classified students in the two most rural classifications as Rural and other students as non-Rural. This designation of rural status was well aligned with that which would be applied by the residents of the state in which University 1 is located. This method is a refinement of an earlier classification of rural/urban status [21] and should be more accurate.

For University 2, gender data were collected from university records. Race/ethnicity data were self-reported using a survey instrument. Again, race/ethnicity was coded dichotomously with White non-Hispanic students coded as White and all other students as non-White.

III. RESULTS

RQ1: Are pretest scores equally correlated with measures of physics achievement for different demographic groups? When pretest scores are used to make academic adjustments early in the semester, the goal is generally to influence some academic outcome variable such as post-test score, test average, or course grade. In these situations, it is important that pretest scores provide equally accurate information for all students. Because some demographic groups in the minority have been shown to have weaker performance post instruction than their majority counterparts, to promote equity, it is particularly important that pretest scores are equally accurate for these groups. Table I presents the mean M and standard deviation of each post-instruction achievement measure as well as the difference in each measure ΔM for each demographic group. In all cases, the group in the minority in the physics classes was compared to the majority group and, therefore, ΔM is positive if the majority group had a higher score than the group in the minority. The significance of the difference was calculated with a t-test and is represented by a superscript on ΔM. The effect size for this difference is measured by Cohen's d. Cohen suggests $d = 0.20$ as a small effect, 0.5 as a medium effect, and 0.80 as a large effect [36]. Table II presents pretest percentage scores as well as the correlation r between the pretest and each achievement measure for each demographic group. The difference of these correlations between groups, Δr, is also presented. The significance of the difference was calculated by bootstrapping with 1000 replications; the significance of the difference is represented by a superscript on Δr. The standard deviations of the bootstrapped correlations were generally small leading to most of the Δr being significantly different at the $p < 0.001$ level. As such, it is more productive to consider the effect size of the difference in correlation. Cohen suggests a correlation of $r = 0.1$ as a small effect, $r = 0.3$ as a medium effect, and $r = 0.5$ as a large effect [36]. This suggests either 0.1, the threshold for small effect, or 0.2, the difference between small and medium effects, could be used as thresholds for a practically meaningful differences in correlation. Sample size affects whether a difference in correlation is significant; however, the very small standard deviations of the correlation coefficients in this study result in differences in correlations of 0.01 or greater being detected as significant suggesting that the small sample sizes of some groups was not an important effect in the analysis.

This work presents the results of 80 t-tests in Tables I and II and, therefore, inflation of Type I error is a concern. A Bonferroni correction, dividing the significance thresholds by the number of statistical tests performed, was applied to all significance tests presented. The corrected thresholds are superscript "a" representing $p < 0.000625$ (originally $p < 0.05$), "b" $p < 0.000125$ (originally $p < 0.01$), and "c" $p < 0.0000125$ (originally $p < 0.001$). The Bonferroni correction is an aggressive error correction method; many additional differences would have been significant or significant at a higher level if it had not been applied. There are many alternate methods to account for error inflation; as such, it is probably more productive to examine the effect sizes of the differences rather than their significance level.

TABLE I. Post instruction achievement measures. Rows are labeled Majority/Minority with N_{MJ} the number of the students in the majority group and N_{MN} the number of students in the group in the minority. M is the mean plus/minus the standard deviation for each group. ΔM is the difference in group means; positive values indicate the majority group with the higher mean. The effect size d for this difference is also reported. Items where the difference in mean represents at least a small effect size have been bolded. The significance of each difference is represented by a superscript. Superscript "a" denotes $p < 0.05$, "b" denotes $p < 0.01$, and "c" denotes $p < 0.001$. A Bonferroni correction was applied to the p values.

			Post-test%				Test Average				Course Grade			
	N_{MJ}	N_{MN}	M_{MJ}	M_{MN}	ΔM	d	M_{MJ}	M_{MN}	ΔM	d	M_{MJ}	M_{MN}	ΔM	d
FMCE-1														
Male/Female	1879	571	$\mathbf{50 \pm 29}$	$\mathbf{38 \pm 24}$	$\mathbf{12.1^c}$	**0.44**	70 ± 15	70 ± 15	0.3	0.02	2.90 ± 1.1	2.97 ± 1.0	-0.07	0.07
White/Non-White	2062	388	$\mathbf{50 \pm 29}$	$\mathbf{40 \pm 26}$	$\mathbf{9.9^c}$	**0.35**	$\mathbf{70 \pm 15}$	$\mathbf{67 \pm 16}$	$\mathbf{3.2^a}$	**0.21**	2.90 ± 1.1	2.73 ± 1.1	0.17	0.16
Non-FGCS/FGCS	2060	390	47 ± 28	44 ± 27	3.4	0.12	$\mathbf{71 \pm 15}$	$\mathbf{68 \pm 15}$	$\mathbf{3.2^a}$	**0.21**	$\mathbf{2.95 \pm 1.0}$	$\mathbf{2.74 \pm 1.1}$	$\mathbf{0.21^a}$	**0.21**
Non-Rural/Rural	1495	133	48 ± 28	53 ± 29	-4.9	0.17	70 ± 15	69 ± 17	0.1	0.01	2.88 ± 1.1	2.83 ± 1.1	0.06	0.05
CSEM-1														
Male/Female	1414	382	$\mathbf{57 \pm 17}$	$\mathbf{52 \pm 17}$	$\mathbf{4.5^c}$	**0.26**	72 ± 15	72 ± 15	-0.4	0.02	2.87 ± 1.0	3.00 ± 1.0	-0.13	0.12
White/Non-White	1484	312	$\mathbf{57 \pm 17}$	$\mathbf{52 \pm 17}$	$\mathbf{4.6^b}$	**0.26**	72 ± 15	70 ± 15	1.7	0.11	2.87 ± 1.0	2.72 ± 1.1	0.15	0.14
Non-FGCS/FGCS	1510	286	56 ± 17	55 ± 18	1.3	0.07	73 ± 15	70 ± 16	2.4	0.16	$\mathbf{2.94 \pm 1.0}$	$\mathbf{2.68 \pm 1.1}$	$\mathbf{0.26^b}$	**0.25**
Non-Rural/Rural	1175	102	56 ± 17	59 ± 17	-2.9	0.17	72 ± 15	72 ± 17	0.8	0.05	2.93 ± 1.1	2.88 ± 1.1	0.04	0.04
FCI-2														
Male/Female	1756	617	$\mathbf{77 \pm 16}$	$\mathbf{69 \pm 16}$	$\mathbf{8.3^c}$	**0.52**	-	-	-	-	3.45 ± 0.66	3.52 ± 0.63	-0.07	0.11
White/Non-White	2036	337	$\mathbf{77 \pm 16}$	$\mathbf{69 \pm 17}$	$\mathbf{8.3^c}$	**0.52**	-	-	-	-	3.45 ± 0.66	3.34 ± 0.69	0.11	0.16
CSEM-2														
Male/Female	1927	610	$\mathbf{65 \pm 16}$	$\mathbf{60 \pm 16}$	$\mathbf{5.5^c}$	**0.35**	76 ± 12	77 ± 13	-0.2	0.02	3.34 ± 0.70	3.43 ± 0.67	-0.09	0.13
White/Non-White	2163	374	$\mathbf{65 \pm 16}$	$\mathbf{58 \pm 17}$	$\mathbf{7.2^c}$	**0.46**	$\mathbf{76 \pm 12}$	$\mathbf{74 \pm 13}$	**2.5**	**0.20**	3.34 ± 0.70	3.30 ± 0.74	0.05	0.06

TABLE II. Pretest scores and the correlation of achievement measures with pretest score for different demographic groups. Rows are labeled Majority/Minority. r is the correlation with pretest score and Δr is the difference in the correlation between the demographic groups. Bolded items represent differences in correlation with $\Delta r > 0.1$. Superscript "a" denotes $p < 0.05$, "b" denotes $p < 0.01$, and "c" denotes $p < 0.001$. A Bonferroni correction was applied to the p values.

	Pretest%				Post-test			Test Average			Course Grade		
	M_{MJ}	M_{MN}	ΔM	d	r_{MJ}	r_{MN}	Δr_{post}	r_{MJ}	r_{MN}	Δr_{test}	r_{MJ}	r_{MN}	Δr_{grade}
FMCE-1													
Male/Female	25 ± 20	19 ± 15	5.6^c	0.29	0.67	0.67	0.00	0.45	0.43	0.02^c	0.29	0.29	0.01^c
White/Non-White	25 ± 20	22 ± 17	3.5^a	0.18	0.67	0.71	-0.04^c	0.45	0.46	-0.01^c	0.29	0.31	-0.02^c
Non-FGCS/FGCS	24 ± 20	21 ± 16	3.1^a	0.16	0.68	0.65	0.03^c	0.45	0.38	0.07^c	0.29	0.23	0.06^c
Non-Rural/Rural	24 ± 19	25 ± 19	-1.2	0.06	0.66	0.72	-0.06^c	0.44	0.41	0.02^c	0.27	0.30	-0.03^c
CSEM-1													
Male/Female	28 ± 11	25 ± 9	3.2^c	0.31	0.44	0.40	0.03^c	0.33	0.28	0.05^c	0.24	0.24	0.00
White/Non-White	28 ± 11	26 ± 10	1.8^a	0.26	0.44	0.38	0.05^c	**0.33**	**0.20**	$\mathbf{0.13^c}$	**0.24**	**0.10**	$\mathbf{0.14^c}$
Non-FGCS/FGCS	27 ± 11	27 ± 10	0.2	0.02	0.44	0.43	0.00	0.33	0.24	0.09^c	0.24	0.18	0.05^c
Non-Rural/Rural	28 ± 10	27 ± 9	1.1	0.11	0.42	0.44	-0.02^c	0.30	0.33	-0.03^c	**0.21**	**0.33**	$\mathbf{-0.12^c}$
FCI-2													
Male/Female	45 ± 18	32 ± 15	13^c	0.76	**0.60**	**0.46**	$\mathbf{0.14^c}$	-	-	-	0.34	0.28	0.06^c
White/Non-White	45 ± 18	36 ± 18	8.8^c	0.48	0.60	0.58	0.01^c	-	-	-	0.34	0.34	-0.00
CSEM-2													
Male/Female	29 ± 11	25 ± 9	4.2^c	0.39	**0.40**	**0.24**	$\mathbf{0.16^c}$	0.29	0.18	0.11^c	0.23	0.19	0.04^c
White/Non-White	29 ± 11	27 ± 11	2.3^a	0.21	0.40	0.42	-0.02^c	0.29	0.29	0.00	0.23	0.26	-0.03^c

There are many significant differences in the achievement measure means between the groups, some with up to a large effect size $d = 0.76$. These samples and the post-test differences have been thoroughly explored in previous works [12, 13, 21]; the mean differences observed in Table I are in agreement with these works. This work does not find a significant post-test difference for rural and non-rural students; this differs from results reported in Henderson, Zabriskie, and Stewart [21] and is due to the more restrictive definition of rural used in this work. The gender and race/ethnicity differences in the post-test scores observed in each sample are not consistently reproduced for test averages or course grades. The correlation differences have not been previously explored. In general, the CSEM pretest was more weakly correlated with post-instruction achievement measures than either the FCI or FMCE pretest. Most Δr in Table II are substantially below the 0.1 threshold, but a number are not; these are highlighted by bolding. Sample CSEM-2 was previously analyzed by Henderson *et al.* [30]; the large difference in the correlation of pretest and post-test between men and women is consistent with their reported difference between men and women on a latent variable measuring conceptual prior knowledge not explained by quantitative performance. This large difference is also present in Sample FCI-2. The gender fairness of the instruments have been examined for these samples [9, 10]; the FCI was shown to be substantially unfair while the CSEM was generally fair. Samples FCI-2 and CSEM-2 have similar student populations which may indicate the source of differences in correlation lie with properties of the student population rather than properties of the instruments. These strong male/female differences are also present to a lesser degree in the test average in CSEM-2, but are not present in course grades. The difference in the correlation with the achievement measures between men and women was not present for any achievement measure in Samples FMCE-1 and CSEM-1 showing these differences are not universal.

Substantial differences in the correlation of pretest score with both test average and course grade were also measured for White and non-White students in Sample CSEM-1. These differences were not observed in Sample FMCE-1 which has a similar student population. This difference suggests the correlation differences do no reside in general features of the populations, but instead, how well the pretest instrument detects details of the differences important to success in the individual classes. A substantial difference in the correlation of pretest scores with test averages was also observed between FGCS and non-FGCS students in Sample CSEM-1; again, this difference was not measured in Sample FMCE-1.

RQ2: Are any differences identified related to the magnitude of the pretest score? To explore Henderson *et al.'s* [30] conjecture that the differences in correlation resulted from differences in pretest scores, linear regression was used. Each pretest score in Table II was used as an independent variable in a regression with Δr as the dependent variable to determine if there was a general relation between the size of the pretest score and the difference in correlation. All variables were normalized. For the post-test, the pretest was not a significant predictor of Δr_{post}; neither was the pretest a significant predictor of Δr_{grade}. For the test average, the pretest was a significant predictor of Δr_{test} [$p = 0.009$, $R^2 = 0.32$]; the regression coefficient for the pretest scores was negative, $\beta_{pre} = -1.3$. This would imply that increasing pretest scores resulted in less consistent correlations between pretest and test average (the intercept is also negative). This is opposite what one would predict if low pretest scores were generating inaccurate measurements of prior preparation. If the difference in Δr resulted from lower pretest scores inaccurately measuring one group, one would expect a similar effect on all post-instruction achievement measures; this was not observed. As such, this work does not support Henderson *et al.'s* [30] conjecture.

IV. DISCUSSION AND CONCLUSION

This work found, in general, pretest scores were consistently correlated with post-test scores, test averages, and course grades for most demographic groups; however, there were notable exceptions. These differences were as large as $\Delta r = 0.16$ between men and women for the CSEM-2 post-test; this difference represents most of the difference between a small and medium effect size. As such, if pretest scores are used to manage instructional interventions designed to improve post-test scores for students in the class from which Sample CSEM-1 was drawn, women will be less accurately assigned to interventions than men. Similar issues with substantial differences in correlation existed by race/ethnicity and first generation status in some samples. In the only substantial difference in correlation detected between rural and non-rural students, the pretest was a more strongly correlated with course grade for rural students. As such, instructors using pretests for decisions with instructional consequences should be aware that they are not equally accurate for all populations. There was little consistency in the groups where large differences in correlation were detected. As Henderson, Stewart, and Traxler suggest [31], as with any instrument, instructors should verify the accuracy of conceptual instruments for their class and student population before using them for decisions that might have consequences for the instructional outcomes of students. Henderson et al.'s [30] conjecture that the differences in correlations were the result of the lower pretest scores of some groups moving their score distribution nearer the pure guessing distribution was not supported by this work. The inconsistent differences in pretest correlations between institutions, examinations, and demographic groups suggest additional research is required to understand these effects. This work was supported by the National Science Foundation under Grant No. EPS-1003907, ECR-1561517, and HRD-1834569.

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

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

[3] 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).

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

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

[6] A. Madsen, S. 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).

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

[8] A. Miyake, L. Kost-Smith, N. Finkelstein, S. Pollock, G. Cohen, and T. Ito, Reducing the gender achievement gap in college science: A classroom study of values affirmation, Science **330**, 1234 (2010).

[9] 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).

[10] R. Henderson, P. Miller, J. Stewart, A. Traxler, and R. Lindell, Item-level gender fairness in the Force and Motion Conceptual Evaluation and the Conceptual Survey of Electricity and Magnetism, Phys. Rev. Phys. Educ. Res. **14**, 020103 (2018).

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

[12] R. Henderson and J. Stewart, Racial and ethnic bias in the Force Concept Inventory, in *AIP conference proceedings* (AIP, 2017) pp. 172–175.

[13] C. Zabriskie, G. Cochran, S. DeVore, R. Henderson, J. Stewart, P. Miller, G. Stewart, and L. Michaluk, The relation of race, ethnicity, and gender to physics conceptual inventory performance (2018), submitted Phys. Rev. Phys. Educ. Res.

[14] E. Brewe, V. Sawtelle, L. Kramer, G. 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).

[15] L. McCullough, An overview on research on gender and underrepresented ethnicities in physics education, in *Getting Started in PER*, Vol. 2 (2018) 4th ed.

[16] The ACT Profile Report - National Graduating Class 2016, ACT Inc., Iowa City, IA (2016).

[17] 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), US Department of Education, National Center for Education Statistics, Institute of Education Sciences, Washington, DC (2016).

[18] 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).

[19] S. 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).

[20] B. Toven-Lindsey, M. Levis-Fitzgerald, P. 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).

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

[22] E. Pascarella, C. Pierson, G. Wolniak, and P. Terenzini, First-generation college students: Additional evidence on college experiences and outcomes, J. High. Educ. **75**, 249 (2004).

[23] E. Cataldi, C. Bennett, and X. Chen, First-generation students: College access, persistence, and postbachelor's outcomes. (NCES 2018-421), US Department of Education, National Center for Education Statistics, Institute of Education Sciences, Washington, DC (2018).

[24] J. Redford and K. Hoyer, First-generation and continuing-generation college students: A comparison of high school and postsecondary experiences. (NCES 2018-009), US Department of Education, National Center for Education Statistics, Institute of Education Sciences, Washington, DC (2018).

[25] X. Chen, STEM attrition: College students' paths into and out of STEM fields. (NCES 2014-001), US Department of Education, National Center for Education Statistics, Institute of Education Sciences, Washington, DC (2013).

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

[27] J. Williams, Cross-national variations in rural mathematics achievement, J. Res. Rural Educ. **20**, 20 (2005).

[28] Advanced Placement Access and Success: How do rural schools stack up?, The College Board, New York, NY (2017).

[29] A. Pierson and H. Hanson, Comparing postsecondary enrollment and persistence among rural and non-rural students in Oregon (REL 2015–076), US Department of Education, Washinton, DC. (2015).

[30] 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).

[31] 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).

[32] US News & World Report: Education, US News and World Report, Washington, DC, https://premium.usnews.com/best-colleges.

[33] A. Traxler, X. 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).

[34] M. Davenport, NCES/CEEB code crosswalk, University of North Carolina - Office of Institutional Research, https://ire.uncg.edu/research/NCES_CEEB_Table/ (2019).

[35] D. Geverdt, Education Demographic and Geographic Estimates Program (EDGE)- Locale Boundaries User's Manual, National Center of Education Statistics, Washington, DC., https://nces.ed.gov/programs/edge/docs/NCES_LOCALE_USERSMANUAL_2016012.pdf.

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