Editors' Suggestion

Demographic gaps or preparation gaps?: The large impact of incoming preparation on performance of students in introductory physics

Shima Salehi,[1,2] Eric Burkholder,[1] G. Peter Lepage,[3] Steven Pollock,[4] and Carl Wieman[1,2]

[1]*Department of Physics, Stanford University, Stanford, California 94305, USA*
[2]*Graduate School of Education, Stanford University, Stanford, California 94305, USA*
[3]*Laboratory for Elementary Particle Physics, Cornell University, Ithaca, New York 14853, USA*
[4]*Department of Physics, University of Colorado Boulder, Boulder, Colorado 80309, USA*

(Received 20 April 2019; published 18 July 2019)

We have studied the impact of incoming preparation and demographic variables on student performance on the final exam in the standard introductory calculus-based mechanics course at three different institutions. Multivariable regression analysis was used to examine the extent to which exam scores can be predicted by a variety of variables that are available to most faculty and departments. The results are surprisingly consistent across the institutions, with only math SAT or ACT scores and concept inventory prescores having predictive power. They explain 20%–30% of the variation in student exam performance in all three cases. In all cases, although there appear to be gaps in exam performance if one considers only demographic variables (gender, underrepresented minority, first generation), once these two proxies of incoming preparation are controlled for, there is no longer a demographic gap. There is only a preparation gap that applies equally across the entire student population. This work shows that to properly understand differences in student performance, it is important to do statistical analyses that take multiple variables into account, covering both subject-specific and general preparation. Course designs and teaching better matched to the incoming student preparation will likely eliminate performance gaps across demographic groups, while also improving the success of all students.

DOI: 10.1103/PhysRevPhysEducRes.15.020114

I. INTRODUCTION

Physics education researchers have made great progress in finding teaching methods that result in improvements in student learning when looking at class averages [1–4]. A recent and growing focus has been to go beyond averages and overall normalized gains to look at how teaching methods impact different students in different ways [5–9]. This is an important step in finding how to best serve the different student subpopulations in our classes, including providing inclusive learning environments for historically underrepresented demographic populations in science, technology, engineering, and math (STEM) fields. That is an essential step for improving the diversity in physics in particular, and STEM fields in general. The first step in such research is to identify which factors are important in determining student outcomes for different populations, and hence, where it would be most effective to focus teaching improvements and research.

We use data from three different institutions to explore the effect of a variety of student characteristics on their score on the final exam in the large introductory calculus-based physics course ("physics 1"). Nearly all prospective engineering students as well as many science students take this course, and students' academic performance in this course is consequential in pursuing STEM majors in their undergraduate studies. This work does not consider all factors that might be important, but rather a set that most physics instructors or departments will have access to, e.g., incoming SAT or ACT scores, demographic information [gender, first generation (FG) status, and underrepresented minority (URM) status], and precourse physics concept inventory (CI) scores.

Of particular concern to many institutions today are the average gaps in performance often seen between different demographic groups, such as course grades, exam scores, and passing rates [10–17]. Underperformance of demographically underrepresented students in physics 1 can have considerable negative influence on their prospect of pursuing STEM fields, thereby preventing the increase of their representation in those fields. The factors that give rise to such gaps and how we can best design learning environments to address them are important unanswered questions. It is important to identify and remove factors that

might produce such gaps, but there is also a danger associated with focusing on such gaps. There are negative consequences to labeling gaps as demographic gaps when the gaps are not arising from demographic status *per se*, but from the factors correlated with it. This mislabeling can result in bias and negative expectations for the labeled demographic group by both instructors and students [18–20].

The most important result of our analysis was that it revealed that differences in math SAT or ACT scores and CI prescores, which we use as admittedly crude proxies of incoming preparation, were sufficient to explain the performance gaps between demographic groups in our data. Thus, it would be misleading and potentially harmful to discuss gaps in performance between females and males, URM and majority students, and first-generation and continuing generation students, as is customarily done, when the differences in performance are not directly arising from causes associated with those distinctions, but rather appear to be due to differences in incoming preparation. The distinctions in performance are between students with good preparation in physics and poor preparation in physics, or more specifically, good math SAT or ACT scores and CI prescores and poor scores on those two measures. This distinction is the same across all demographic groups.

Addressing a range of incoming preparation is a challenge faced by every physics instructor. How can an instructor best address the range of students in their class in their instruction? Since no institution could (or should) base its admissions entirely on physics preparation, there is an inherent range of physics preparation in every class. Also, when a student does poorly, how much of that is due to weaknesses in their preparation relative to other students, how much is the result of instruction, and how much is due to other possible factors, such as student demographics and the relationship to social-psychological issues that an instructor may or may not be able to affect?

This paper is following the increasing, but still relatively new, trend to explore questions about factors correlated with student performance using more extensive statistical analyses such as multiple regressions and structural equation modeling (SEM) [21–28]. In an introductory physics course many factors can contribute to student performance, and some of these factors may not act independently—they may interact in complex ways. This can only be explored using multivariable regression analysis. Furthermore, SEM can provide additional information by testing for potential structural relations, such as mediation pathways, between multiple different factors.

The research questions to be explored in this paper are largely empirical.

1. How much of the variation in performance in the standard introductory calculus-based college physics course (physics 1) at three institutions can be explained by readily obtained measures of incoming student characteristics?
2. What are some underlying mechanisms for gender, FG, and URM academic performance gaps in physics 1 and do those justify the singling out of these particular gaps?
3. How similar are the answers to (1) and (2) across different institutions? (The limited sample available for this work allows only a start at answering this question.)

We are not providing answers to these questions that apply to all populations of college students, but rather preliminary observations that we hope will stimulate others to carry out similar analyses so that a larger body of data spanning more institutions and student populations can be accumulated to provide more generalizable answers. Such data are necessary to address the more fundamental question that we and many others find particularly important, namely, what forms of instruction are the most effective at achieving success for the maximum number of students in our courses, given the inherent differences present in any student population?

II. METHODS

We looked at the large physics 1 course at three large, research-intensive institutions: a highly selective east coast university (HSEC), a highly selective west coast university (HSWC), and a large public research university in the middle of the country (PM). Physics 1 is the standard introductory course that is offered by most physics departments and is taken primarily by students intending to major in engineering, as well as some chemistry and physics majors and some premedical students. In Table I we list some characteristics of the institutions and the students in physics 1 at the three institutions.

The data we have for all three institutions are gender, URM, and FG status, proxies for students' incoming preparation (their pre- and postcourse concept inventory test scores, a mix of math SAT or ACT scores), and their course performance (physics 1 final exam scores). Both HSWC and PM used the Force and Motion Conceptual Evaluation (FMCE), while HSEC used the Force Concept Inventory (FCI) as a physics concept inventory [29,30]. Students were considered URM if they were nonwhite, non-Asians, and were considered first-generation college attending if neither of their parents had a four-year college degree. As we had a mixture of ACT and SAT scores, we converted all of them to percentile scores using available conversion tables, and used the resulting percentile scores in our regression models [31].

In addition, we have particular pieces of data for only one or two institutions. We included these in our model analyses for those institutions to test for the importance. These data include taking a supplementary help session targeting students with weaker preparation (HSWC, 2018), the number that had taken Advanced Placement (AP) physics (HSWC and HSEC), and midterm exam scores

TABLE I. Institutional characteristics and characteristics of students in physics 1.

Institutional characteristics	HSEC	HSWC	PM
No. students per year taking physics 1	194 (2012), 185 (2013)	466 (2017), 518 (2018)	4 offerings 2015–2017, ∼1100 per class
Math SAT top 25th and 75th percentile	790, 700	800, 730	690, 570
% in top 10% of HS class	86	96	29
Physics 1 class characteristics			
Average percentile math SAT or ACT score	97	97	89
Average prescore on concept inventory (%)	63, 61	58, 53	38–49
Normalized pre-post gain on CI	0.40, 0.36	0.44, 0.47	0.49–0.54

(HSWC, 2017). Although we have the physics 1 course grades for all the institutions, we only used final exam scores in our analysis because the grading standards and the course components that go into the calculation of these course grades varied greatly across the three institutions. The structure, administration, and grading of the final exams were similar.

We carried out multivariable linear regression analyses for each of the three institutions in the dataset, using gender, FG, and URM status, as well as various measures of incoming preparation to predict the final exam score. We normalized all the continuous variables in these analyses in terms of the sample standard deviation ("z scores"), so the coefficients in the models can be directly interpreted as the fraction of a standard deviation in the outcome variable for a 1 standard deviation change in the continuous predictive variable. For the categorical variables, such as gender, the coefficient refers to the effect of changing from 0 to 1. In these analyses, we examined which combination of the aforementioned variables would provide the simplest, best-fitting model to predict students' final exam scores.

In Appendix A, we provide more details as to how the models are evaluated and the criteria used to include terms to find the simplest, best-fitting model. In the model evaluation, we focused on the value of R-squared [(explained variation of the outcome variable)/(total variation of the outcome variable)] (the larger the R-squared, the better the model fit), and the value of the Akaike information criterion (AIC) [32]. The AIC is a standard criterion for evaluating the quality of a predictive model that takes into account the parsimony of the model, so it considers the number of variables in the model as well as its predictive power (the smaller the AIC score, the better the model).

With regression analysis, one can explore which and to what extent predictors correlate with the dependent variable of interest, but cannot directly explore the relationship between the predictors themselves. To determine these relationships, and the corresponding effects on the dependent variable, student exam performance, we employed structural equation modeling (SEM; see Appendix B) using the LAVAAN package in R [33,34]. In the structural equation modeling, we tested whether incoming preparation is a mediator for the effect of demographic status on student performance. Note that SEM does *not* test for causality—only randomized, controlled experiments can be used to test causality—rather it tests for structural relationships in the data. In other words, we tested whether students from different demographic status have different levels of incoming preparation, and how these differences lead to differences in exam scores. For a primer on SEM, we refer the interested reader to Ref. [35].

III. RESULTS

A. Predictors of exam performance:

If one looks only at average exam scores for the various demographic groups, there are significant differences as shown in Table II. This analysis gives an estimate of demographic gaps without controlling for incoming preparation of students from different demographic groups. However, when using multivariable regression to control for students' incoming preparation as measured by CI prescores and math SAT or ACT scores, the direct effects of demographic variables on student outcomes become insignificant. The coefficients of demographic status in this regression analysis give an estimate of demographic gaps when controlling for incoming preparation as measured by math SAT or ACT and CI prescore. We illustrate this explicitly in Fig. 1, as well as Table II—the data presented in Fig. 1 and Table II are identical, and the numbers in Table II are provided for the reader interested in the details of effect sizes and R-squared values. The blue (leftmost) columns in Fig. 1 show the coefficient of the demographic status for a model that predicts final exam scores including only the single demographic variable (equivalent to a simple t-test); these correspond to the top row of parts (a), (b), and (c) of Table II. The teal (center) columns then show the size of this coefficient in the model when math SAT or ACT scores are added to the model [second row of parts (a), (b), and (c) in Table II], and finally, the yellow

TABLE II. Various regression models comparing the effects of incoming preparation and (a) gender, (b) URM status, or (c) FG status on final exam across the different institutions. The regression coefficients are normalized such that they may be interpreted as an effect size. $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$ with no correction to p values for multiple comparisons. As this represents nine regression analyses performed on each of the three datasets, correcting the p values for multiple comparisons makes the coefficient associated with PM FG status statistically consistent with zero. The p values for b_{Math} and b_{CI} are so small that the significance of these coefficients is not affected by correction for multiple comparisons. All values of R^2 shown are adjusted R^2 values.

(a) Predictor of final exam score	HSEC	HSWC	PM
Gender	$b_{\text{gender}} = -0.24$ (0.10)**	$b_{\text{gender}} = -0.26$ (0.07)***	$b_{\text{gender}} = -0.28$ (0.04)***
	$R^2 = 0.01$	$R^2 = 0.02$	$R^2 = 0.02$
Math SAT/ACT + Gender	$b_{\text{Math}} = 0.28$ (0.05)***	$b_{\text{Math}} = 0.37$ (0.03)***	$b_{\text{Math}} = 0.37$ (0.02)***
	$b_{\text{gender}} = -0.26$ (0.10)***	$b_{\text{gender}} = -0.22$ (0.07)***	$b_{\text{gender}} = -0.23$ (0.04)***
	$R^2 = 0.09$	$R^2 = 0.15$	$R^2 = 0.16$
Math SAT/ACT + CI + Gender	$b_{\text{CI}} = 0.34$ (0.05)***	$b_{\text{CI}} = 0.39$ (0.03)***	$b_{\text{CI}} = 0.38$ (0.02)***
	$b_{\text{Math}} = 0.2$ (0.07)***	$b_{\text{Math}} = 0.22$ (0.03)***	$b_{\text{Math}} = 0.26$ (0.02)***
	$b_{\text{gender}} = -0.04$ (0.10)	$b_{\text{gender}} = -0.04$ (0.06)	$b_{\text{gender}} = -0.02$ (0.04)
	$R^2 = 0.18$	$R^2 = 0.27$	$R^2 = 0.28$

(b) Predictor of final exam score	HSEC 11–13	HSWC 17–18	PM 14–17
URM	$b_{\text{URM}} = -0.51$ (0.13)***	$b_{\text{URM}} = -0.38$ (0.11)***	$b_{\text{URM}} = -0.16$ (0.05)***
	$R^2 = 0.03$	$R^2 = 0.03$	$R^2 = 0.004$
Math SAT or ACT + URM	$b_{\text{Math}} = 0.24$ (0.06)***	$b_{\text{Math}} = 0.46$ (0.05)***	$b_{\text{Math}} = 0.38$ (0.02)***
	$b_{\text{URM}} = -0.18$ (0.16)	$b_{\text{URM}} = -0.05$ (0.10)	$b_{\text{URM}} = -0.002$ (0.04)
	$R^2 = 0.07$	$R^2 = 0.22$	$R^2 = 0.15$
Math SAT or ACT + CI + URM	$b_{\text{CI}} = 0.34$ (0.05)***	$b_{\text{CI}} = 0.37$ (0.05)***	$b_{\text{CI}} = 0.38$ (0.02)***
	$b_{\text{Math}} = 0.16$ (0.06)**	$b_{\text{Math}} = 0.31$ (0.05)***	$b_{\text{Math}} = 0.26$ (0.02)***
	$b_{\text{URM}} = -0.16$ (0.15)	$b_{\text{URM}} = -0.02$ (0.10)	$b_{\text{URM}} = -0.02$ (0.04)
	$R^2 = 0.18$	$R^2 = 0.33$	$R^2 = 0.28$

(c) Predictor of final exam score	HSEC	HSWC	PM
FG	$b_{\text{FG}} = -0.24$ (0.22)	$b_{\text{FG}} = -0.53$ (0.13)***	$b_{\text{FG}} = -0.38$ (0.05)***
	$R^2 = 0.0005$	$R^2 = 0.04$	$R^2 = 0.02$
Math SAT or ACT + FG	$b_{\text{Math}} = 0.27$ (0.05)***	$b_{\text{Math}} = 0.45$ (0.05)***	$b_{\text{Math}} = 0.37$ (0.02)***
	$b_{\text{FG}} = -0.15$ (0.22)	$b_{\text{FG}} = -0.16$ (0.12)	$b_{\text{FG}} = -0.17$ (0.05)
	$R^2 = 0.07$	$R^2 = 0.22$	$R^2 = 0.15$
Math SAT or ACT + CI + FG	$b_{\text{CI}} = 0.34$ (0.05)***	$b_{\text{CI}} = 0.37$ (0.05)***	$b_{\text{CI}} = 0.38$ (0.02)***
	$b_{\text{Math}} = 0.19$ (0.05)***	$b_{\text{Math}} = 0.30$ (0.05)***	$b_{\text{Math}} = 0.25$ (0.02)***
	$b_{\text{FG}} = -0.12$ (0.21)	$b_{\text{FG}} = -0.12$ (0.11)	$b_{\text{FG}} = -0.11$ (0.04)*
	$R^2 = 0.18$	$R^2 = 0.33$	$R^2 = 0.28$

(rightmost) columns show the size of the demographic coefficient after the CI prescores (indicating subject-specific preparation) are also added to the model as well as math SAT or ACT [third row of parts (a), (b), and (c) of Table II].

As the gender (leftmost) panel of Fig. 1 shows, the gender gap, i.e., the gender coefficient, changes little when math SAT is added to the regression model (the change from the blue to teal bar in the gender panel), reflecting the fact that there is very little difference between average male and female math SAT or ACT scores. However, there is a large change in the gender coefficient when the CI prescore is added to the regression model (the change from the teal to yellow bar in the gender panel). This change implies a significant average difference in the CI prescores between males and females. For the URM gap, a different pattern is apparent. The size of the URM coefficient in the regression model shows more initial variation across institutions, but for all three institutions, when math SAT or ACT score is added to the model, the URM gap is drastically reduced and

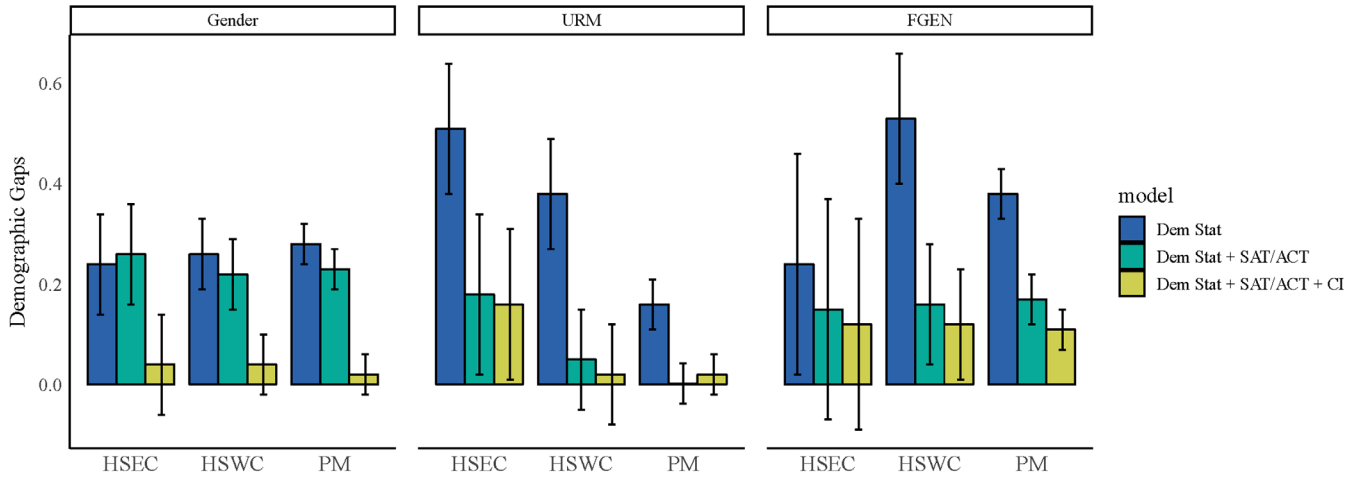

FIG. 1. Size of coefficient for the demographic status as predicted by regression models for each institution. First model (blue) shows the coefficient where only the respective demographic status (gender, URM, FG) is included in the model; second model (teal) shows the demographic coefficient when math SAT or ACT score is added as a predictor; and the third model (yellow) is the coefficient when the CI prescore is added as a predictor as well as math SAT or ACT score. The error bars represent the standard error of the coefficients. As shown in Table II, the regression models with only demographic status have R-squared values of 0.03 or less, but these increase to 0.2–0.3 when measures of incoming preparation are added to the model.

becomes insignificant. For the FG gap, there is less obvious consistency across institutions, except that adding math SAT or ACT score to the model changes the FG gap significantly. Overall, the majority of URM and FG gaps could be explained by math SAT or ACT score, while math SAT or ACT was negligible in explaining the gender gap. However, all of the gender gap could be explained by the CI prescore. This observation that these two incoming preparation measures had different explanatory power for different demographic gaps demonstrates the importance of having multiple measures to adequately characterize students' incoming preparation, particularly including subject-specific measures. Table III shows the regression models with only incoming preparation variables. As shown, the CI and math SAT or ACT can predict almost one-third of the overall variance in the final exam scores. Furthermore, adding demographic variables makes a negligible improvement to the adjusted R-squared of the model.

In summary, while demographic performance gaps on the physics 1 final exam exist at these institutions, the gaps can be explained by differences in students' incoming preparation as estimated by the two measures of math SAT or ACT and CI prescore. These two measures are actually rather crude measures for incoming preparation in physics; therefore, it is striking that they are sufficient to eliminate the significance of the demographic variables.

In Appendix B, we use SEM to examine in more detail how CI prescores and math SAT or ACT scores mediate the effect of demographic characteristics on final exam scores. This quantifies how the different demographic gaps are mediated by math SAT or ACT score and CI prescore. The SEM confirms that demographic gaps in final exam scores are mediated by incoming preparation. This is the same qualitative information presented in Table II and Fig. 1 but provides additional quantitative statistical tests [35].

We have also looked at the possible contributions of several other variables, many of which we had for only a subset of institutions. For HSWC (2017), composite SAT or ACT score had less predictive power than math SAT or ACT score: the AIC of the model with only math SAT or ACT as predictor was lower than the model with only composite SAT or ACT as a predictor (820 as opposed to 837). Also, addition of the composite SAT or ACT to the model with math SAT or ACT and CI prescore as predictors did not change the R^2 of the model (0.30 for both models). For HSWC (2018), we were able to examine the effect of cumulative student university GPA at the start of the course.

TABLE III. Models predicting the final exam scores only by the two measures of incoming preparation across the three different institutions and multiple years at each institution. The regression coefficients are normalized such that they may be interpreted as an effect size.

Predictor of final exam score	HSEC	HSWC	PM
Math SAT or ACT percentile plus pre CI	$b_{CI} = 0.34$ (0.05)*** $b_{Math} = 0.20$ (0.05)*** $R^2 = 0.18$	$b_{CI} = 0.34$ (0.05)*** $b_{Math} = 0.35$ (0.05)*** $R^2 = 0.34$	$b_{CI} = 0.38$ (0.02)*** $b_{Math} = 0.26$ (0.02)*** $R^2 = 0.28$

This effect is significant when adding GPA to the model in Table III—R^2 goes from 0.34 to 0.47—an expected result as this captures other factors of students' adjustment to the college academic environment. We also examined the effect of a supplementary weekly help session (HSWC 2018) and found no significant impact on final exam score. The HSWC (2017) student scores on a set of Colorado Learning Attitudes about Science Survey (CLASS) [36] questions reporting on their self-efficacy, both pre- and postcourse, were also found to be negligible predictors. In previous work [20], we had already seen that these were the only items on the CLASS survey that showed significant variation across these populations, but here we see that variation is not correlated with exam performance. We also considered the interaction between demographic variables and the factors that we expected might be significant—for example, if the supplementary weekly help session disproportionately helped URM students. No such two-way interactions were found to be significant, and it was assumed therefore that no higher-order interactions would contribute significantly. For HSEC and PM, we had several years of equivalent data, and we found that the inclusion of a random effect of "year" in the model was not significant—these findings are consistent over time. The consistency at PM was probably because they had a standard body of exam questions from which questions were chosen. Multiple instructors, including some who had taught the course in previous years, reviewed the exam before administering it to ensure equivalence between the different years. At HSEC, the instructor was the same across years, which likely contributes to this consistency. HSWC exams have both open-ended and multiple-choice questions. PM exams have only multiple-choice questions, with more emphasis on conceptual understanding. HSEC exams were a combination of short answer and multiple-choice questions, which were largely conceptual, and longer free-response questions involving calculations. A previous study has shown that student responses to multiple-choice and free-response exam questions in introductory physics are essentially equivalent, so we do not expect the precise structure of the exam to have a significant impact on the results [37].

We also used regression models to predict the CI postcourse scores, as performance gaps on such tests has been a topic of interest [13]. For all three institutions, the distribution of CI postscores is highly distorted, showing a strong ceiling effect. With such distorted distributions, it is questionable as to how valid regression models will be. Similar to what was reported by Day *et al.*, we found that different types of statistical analyses suggest different conclusions [38,39]. A linear regression model indicates that some demographic variables are statistically significant while others are not, but if we model the natural log transform of CI postscore to address the ceiling effect, the model indicates that none of the demographic variables are significant. We interpret this to mean that the CI post distribution is sufficiently distorted that one cannot obtain statistically reliable results from such analyses. For that reason, we present no analysis and make no claims concerning the CI postscores. We do note that adding the CI postscore as a predictor in our linear regression model of the final exam improves the value of R-squared by about 0.1 for all three institutions. This does imply that there is considerable overlap in what the CI and the final exams are measuring, even though they appear to have little resemblance.

B. Failure analysis:

Our regression analysis (Table III) implies that the variations in students' incoming preparation account for 20%–30% of the variation in final exam scores. These R-squared values may seem modest to some, but they have career-altering implications for students who are poorly prepared, which we illustrate with a "failure analysis." This analysis compares the probability of being in the bottom quartile of final exam scores ("failure") for the top and bottom quartiles of incoming preparation. As we show in Appendix C, the probability of failure can be calculated from the value of R-squared found in the regression analysis reported in Table III.

The striking results of this analysis are shown in Fig 2. For an R-squared of 0.34 (that of HSWC), this shows that a student who comes in with preparation in the bottom quartile has about a factor of 4 higher probability of being in the bottom quartile of the grade distribution than a student who starts the course in the upper quartile of preparation. If one considers bottom quartile exam scores as failing, this means that poorly prepared students are 4 times more likely to fail their physics 1 final exams than peers with good incoming preparation.

As discussed in Appendix C, the calculation of the failure ratio purely in terms of R-squared assumes that the incoming preparation variables have a normal

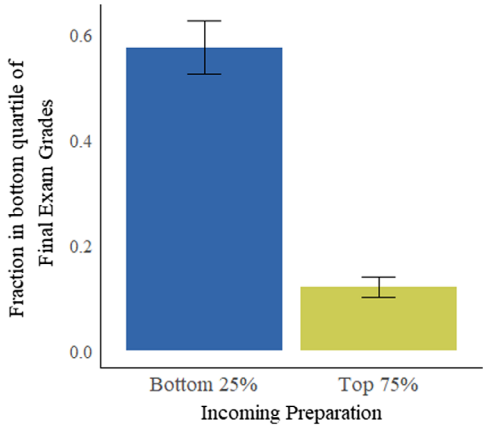

FIG. 2. Probability of a student scoring in the bottom 25% of the class as a function of their preparation as measured by the weighted sum of their math SAT or ACT scores and CI prescores. Error bars indicate the standard error of the measurement.

distribution. However, we have also calculated the failure ratio using a full logistic regression which is not sensitive to these distributions, and we get essentially the same value, indicating that the calculation using just R-squared is quite accurate. Figure 5 in Appendix C allows one to estimate the corresponding failure ratio for any value of R-squared, including the other two institutions discussed here.

IV. DISCUSSION

The first notable result of this analysis is the degree of similarity across three different institutions in spite of having different admissions criteria and selectivity and locally defined physics courses and final exams. Comparing HSWC and PM, we see substantial differences in both average math SAT or ACT and CI prescores, but they have very nearly identical predictive power. The variation in final exam scores predicted by these two factors is about 30% in both cases. Although HSEC is using a different CI, the FCI rather than the FMCE, the final model looks fairly similar to the others, with the same coefficient for the CI term and a 30% smaller coefficient for math SAT or ACT, and a somewhat smaller value of R-squared (0.2 rather than 0.3).

It is natural to wonder, particularly given the similarity of the best predictive models across institutions, how similar the exams and the teaching methods in use at these different institutions are. A detailed analysis is beyond the scope of this paper, but we can provide some general observations. Looking at the final exams, they appear rather similar, with the PM exam being slightly easier in terms of complex quantitative calculations and having somewhat more emphasis on basic concepts. In terms of teaching methods, the HSEC course was largely traditional in all aspects (teaching methods have since been modified), while PM was quite interactive. Peer instruction was used extensively in lectures, and recitation sections used Tutorials in Introductory Physics [40] or similar active learning approaches. HSWC is in between, with similar activities in section to PM, and limited use of peer instruction in lectures.

The second notable, and arguably most important, result in this paper is what it says about the gaps in performance associated with demographic characteristics. It shows that it is misleading to do simple *t*-test comparisons of different demographic groups such as male-female or URM and majority students. To properly understand the variations in student performance across different demographic groups, it is necessary to do regression analyses taking into account incoming preparation, and those measures of incoming preparation need to include both general levels of preparation and subject-specific measures. Such a regression analysis provides an entirely different picture from the *t*-test. For example, across all three institutions, the initial size of the gender gap as predicted by a single-variable model was very similar, about 0.2 standard deviations. When we controlled for students' general incoming preparation as measured by math SAT or ACT score, there was little change in the gap. However, when we also controlled for a subject-specific measure of incoming measure, CI prescore, the gender gap became insignificant for all three institutions. This implies that for these three institutions, once one takes into account differences in students' physics-specific incoming preparation, there is no statistically significant gender gap. For URM gaps, the size of the gap as predicted by a single-variable model varied substantially across institutions. However, when we controlled for students' math SAT or ACT score, the gap became insignificant for all three universities. For FG gaps, the respective sizes of the gaps were quite different at each institution, but controlling for math SAT or ACT scores nearly eliminated the gaps and the differences between institutions.

To emphasize, there are small—if any—gaps in performance associated with demographic differences. There are only performance differences associated with differences in incoming preparation as measured by two proxies: math SAT or ACT and CI prescore. It is notable that math SAT or ACT and CI scores are two rather crude proxies for incoming preparation. The math SAT covers a variety of math knowledge, but little if any of that seems to be an important component to students' success in physics 1 for many of these students. We have explored students' use of math in physics 1 at HSWC in particular, and there was no indication that their performance was limited by math skills. They have mastered all the math they need in physics 1, although their application of math to physical situations is often weak. That is not tested in the math SAT, however, so we attribute the significance of the math SAT or ACT score not to math *per se*, but rather how this score represents some broader level of math-science preparation.

The FCI and FMCE probe a very limited aspect of physics mastery that is needed in physics 1. They test mastery of a limited set of the physics concepts covered, and they are entirely nonmathematical, so they probe nothing about quantitative reasoning and calculational skills which are used extensively on the final exams. Nevertheless, this work shows that it is important to have both general and physics-specific measures of incoming preparation, and that rather crude proxies for each of these is sufficient to explain the apparent demographic differences in performance. We cannot identify what factors are important in determining the level of incoming preparation. We initially expected that it would be differences in what high school physics courses were taken, but we analyzed that for HSWC, and we found that all demographic groups at this institution had the same distribution of taking AP physics, regular high school physics, and no physics, even though the groups had different average CI prescores and math SAT or ACT scores.

Other analyses that looked at gender differences in physics courses without considering incoming preparation have been used to argue for the importance of social-psychological effects, such as stereotype threat, on

the performance of women students [13]. Our analysis, however, shows one can explain all the gender gaps by just performance on a low-stake concept inventory without including any social-psychological factors, at least for the student populations we have considered. This is consistent with the findings of Kost *et al.* [25,26]. This conclusion is also supported by the lack of a correlation between final exam scores and our attitudinal measures of self-efficacy, and the fact that gender did not moderate the correlation between low-stake concept inventory and high-stake final exam performances. However, we cannot rule out some contribution from social-psychological factors, such as test anxiety or stereotype threat. These factors may have affected performance for both the SAT and final exam (they are both high-stakes assessments). Further investigations of this are needed.

V. CONCLUSION

We have examined the variations in the final exam scores in physics 1 across three institutions. This course is a prerequisite for many engineering and science fields, and therefore demographic performance gaps in the course could be consequential in perpetuating the underrepresentation of some demographic groups in STEM fields. We observed significant demographic gaps in final exam scores for all three institutions. However, when we controlled for students' incoming preparation, in all cases the gaps became insignificant or drastically reduced in size. We find that only incoming math SAT or ACT scores and concept inventory prescores together predict 20%–30% of the variation in final exam scores. This is surprisingly consistent across three rather different institutions. Similar analysis from a broader range of institutions is needed to determine the generality of these observations. This will allow further studies on the extent to which different teaching methods might reduce the effects of differences in preparation.

The fraction of the variance explained by the two measures of preparation we have used is substantial, but much less than 1, indicating that there are other important variables in student success. Some students with apparently weak preparation still do quite well. We are carrying out further studies to find out what are these important "hidden variables" that determine the rest of the variance. It should also be noted that the analysis in this paper is correlational. While it is plausible that weak preparation causes low exam performance, this does not demonstrate that. It is possible that there is some unmeasured factor (e.g., test anxiety) that causes both lower scores on our measures of incoming preparation and lower final exam performance.

We hope that the analysis presented here will stimulate others to collect and publish similar results to provide a baseline to better understand the factors contributing to student performance. Understanding these factors can further help us design instructional practices that will benefit more students, including underrepresented minorities. This work shows that incoming preparation is a major predictor for student performance in physics 1, and when controlling for incoming preparation, there remain no demographic performance gaps. Therefore, if we want to improve the outcomes of students from different demographic groups, we have to better address the variation in incoming preparation for all students. This work shows that creating instruction that enhances the success of every student across the full range of incoming preparations is also the solution to eliminating gaps in the performance across demographic groups.

Future work will determine how to best do this, but we can offer some potential suggestions. Better matching the introductory course to the range of background preparations of the student population would likely ensure that many more students, particularly those who had the misfortune to attend K–12 schools that provided weaker education in STEM in general and physics in particular, would achieve better outcomes.

In our brief examination of the physics 1 course at the three institutions, it appears that the level and pace of the course is primarily targeted towards the better-prepared students in the distribution, making the course particularly challenging for students with less preparation, and hence, their results are more sensitive to their preparation level. It is plausible that adjusting the course level to better match the preparation of the less prepared students would improve their performance and reduce the sensitivity to preparation, while having a very small impact on the learning of the best prepared. Another option would be to provide greater resources in the teaching of the course, such as classes with more instructor time, or adding courses to the sequence to provide a greater range of students the opportunity to start with a course matched to their preparation. Of course, that would require additional resources, but in these institutions and a number of others we have examined, the amount of resources expended per credit hour in the science disciplines is far greater for the upper level students than for lower level students such as considered here. If it is an institutional priority to maximize the diversity of a student body that is successfully pursuing STEM careers, reversing that inequality in the expenditure of educational resources between upper and lower level courses would very likely help.

ACKNOWLEDGMENTS

The authors would like to acknowledge the students and instructors from the courses studied, as well as institutional research helpers for providing access to data.

APPENDIX A: MULTIPLE REGRESSION ANALYSIS & MODEL EVALUATION

For model evaluation, one should look for the simplest best-fitting model: the model that has the best fit with the least number of variables (parsimony). One should add

TABLE IV. Different regression models fitted to HSEC data to predict students' final exam score. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	Model 11	Model 12	Model 13	Model 14
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Gender (female = 1, male = 0)	−0.242 (0.103)**			−0.261 (0.099)***	0.006 (0.101)		−0.038 (0.100)	−0.038 (0.100)	−0.036 (0.100)	−0.034 (0.100))	−0.036 (0.100)	−0.034 (0.100)	−0.032 (0.100)	−0.024 (0.103)
Math SAT or ACT		0.275 (0.050)***		0.279 (0.050)***		0.196 (0.048)***	0.198 (0.049)***	0.190 (0.065)***	0.199 (0.049)***	0.229 (0.054)***	0.194 (0.066)***	0.198 (0.065)***	0.231 (0.054)***	0.203 (0.066)***
Pretest CI			0.385 (0.048)***		0.386 (0.050)***	0.342 (0.048)***	0.336 (0.051)***	0.336 (0.051)***	0.317 (0.069)***	0.335 (0.051)***	0.319 (0.071)***	0.337 (0.051)***	0.313 (0.069)***	0.325 (0.071)***
Gender * math SAT or ACT								0.019 (0.095)			0.010 (0.098)	0.090 (0.105)		0.066 (0.119)
Gender * pretest CI									0.039 (0.099)		0.036 (0.103)		0.046 (0.099)	0.027 (0.103)
Math SAT or ACT * pretest CI										0.059 (0.042)		0.076 (0.046)	0.060 (0.042)	0.094 (0.070)
Gender * math SAT or ACT * pretest CI														−0.034 (0.094)
Constant	0.110 (0.069)	−0.002 (0.050)	−0.0004 (0.048)	0.117* (0.067)	−0.003 (0.066)	−0.002 (0.047)	0.015 (0.065)	0.014 (0.065)	0.020 (0.067)	−0.0003 (0.066)	0.020 (0.067)	−0.006 (0.066)	0.006 (0.068)	−0.007 (0.070)
AIC	1072.2	1046.9	1016.8	1042	1018.8	1000.9	1002.7	1004.7	1004.6	1002.7	1006.6	1004	1004.5	1007.8
Observations	378	377	378	377	378	377	377	377	377	377	377	377	377	377
R^2	0.015	0.074	0.149	0.091	0.149	0.185	0.185	0.185	0.185	0.189	0.185	0.191	0.190	0.191
Adjusted R^2	0.012	0.071	0.147	0.086	0.144	0.180	0.178	0.176	0.177	0.181	0.174	0.180	0.179	0.176
Residual standard error	0.994 (d.o.f. = 376)	0.965 (d.o.f. = 375)	0.924 (d.o.f. = 376)	0.957 (d.o.f. = 374)	0.925 (d.o.f. = 375)	0.907 (d.o.f. = 374)	0.908 (d.o.f. = 373)	0.909 (d.o.f. = 372)	0.909 (d.o.f. = 372)	0.906 (d.o.f. = 372)	0.910 (d.o.f. = 371)	0.907 (d.o.f. = 371)	0.907 (d.o.f. = 371)	0.909 (d.o.f. = 369)
F statistic	5.557 (d.o.f. = 1; 376)**	29.919 (d.o.f. = 1; 375)***	65.765 (d.o.f. = 1; 376)***	18.666 (d.o.f. = 2; 374)***	32.797 (d.o.f. = 2; 375)***	42.360 (d.o.f. = 2; 374)***	28.223 (d.o.f. = 3; 373)***	21.122 (d.o.f. = 4; 372)***	21.157 (d.o.f. = 4; 372)***	21.717 (d.o.f. = 4; 372)***	16.883 (d.o.f. = 5; 371)***	17.509 (d.o.f. = 5; 371)***	17.381 (d.o.f. = 5; 371)***	12.475 (d.o.f. = 7; 369)***

TABLE V. Different regression models fitted to HSWC data to predict students' final exam score. * $p < 0.05$, ** $p < 0.01$, ***$p < 0.001$.

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	Model 11	Model 12	Model 13	Model 14
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Gender (female = 1, male = 0)	−0.257 (0.071)***			−0.217 (0.066)***	−0.013 (0.065)		−0.035 (0.063)	−0.035 (0.063)	−0.031 (0.063)	−0.033 (0.063)	−0.032 (0.063)	−0.033 (0.063)	−0.029 (0.063)	0.001 (0.071)
Math SAT or ACT		0.375 (0.033)***		0.369 (0.033)***		0.220 (0.033)***	0.221 (0.033)***	0.186 (0.043)***	0.220 (0.033)***	0.186 (0.040)***	0.197 (0.044)***	0.160 (0.047)***	0.184 (0.040)***	0.184 (0.050)***
Pretest CI			0.480 (0.031)***		0.478 (0.032)***	0.393 (0.033)***	0.388 (0.034)***	0.387 (0.034)***	0.347 (0.044)***	0.400 (0.035)***	0.355 (0.046)***	0.397 (0.035)***	0.358 (0.045)***	0.359 (0.046)***
Gender * math SAT or ACT								0.080 (0.061)			0.052 (0.067)	0.067 (0.062)		−0.016 (0.086)
Gender * pretest CI									0.095 (0.063)		0.073 (0.069)		0.094 (0.063)	0.096 (0.072)
Math SAT or ACT * pretest CI										−0.058 (0.037)		−0.051 (0.038)	−0.057 (0.037)	−0.027 (0.047)
Gender * math SAT or ACT * pretest CI														−0.075 (0.078)
Constant	0.123 (0.049)**	0.000 (0.033)	0.000 (0.031)	0.104 (0.046)**	0.006 (0.044)	0.000 (0.030)	0.017 (0.043)	0.019 (0.043)	0.027 (0.044)	0.038 (0.045)	0.026 (0.044)	0.038 (0.045)	0.049 (0.046)	0.036 (0.047)
AIC	2222.5	2116.3	2029.7	2107.5	2031.6	1988.6	1990.3	1990.6	1990.1	1989.9	1991.5	1990.7	1989.7	1992.5
Observations	786	786	786	786	786	786	786	786	786	786	786	786	786	786
R^2	0.017	0.141	0.230	0.153	0.231	0.271	0.272	0.273	0.274	0.274	0.274	0.275	0.276	0.277
Adjusted R^2	0.015	0.140	0.229	0.150	0.229	0.270	0.269	0.270	0.270	0.270	0.270	0.270	0.271	0.271
Residual standard error	0.992 (d.o.f. = 784)	0.928 (d.o.f. = 784)	0.878 (d.o.f. = 784)	0.922 (d.o.f. = 783)	0.878 (d.o.f. = 783)	0.855 (d.o.f. = 783)	0.855 (d.o.f. = 782)	0.855 (d.o.f. = 781)	0.854 (d.o.f. = 781)	0.854 (d.o.f. = 781)	0.855 (d.o.f. = 780)	0.854 (d.o.f. = 780)	0.854 (d.o.f. = 780)	0.854 (d.o.f. = 778)
F statistic	13.181 (d.o.f. = 1; 784)***	128.419 (d.o.f. = 1; 784)***	234.799 (d.o.f. = 1; 784)***	70.449 (d.o.f. = 2; 783)***	117.276 (d.o.f. = 2; 783)***	145.886 (d.o.f. = 2; 783)***	97.274 (d.o.f. = 3; 782)***	73.444 (d.o.f. = 4; 781)***	73.624 (d.o.f. = 4; 781)***	73.691 (d.o.f. = 4; 781)***	58.987 (d.o.f. = 5; 780)***	59.195 (d.o.f. = 5; 780)***	59.485 (d.o.f. = 5; 780)***	42.612 (d.o.f. = 7; 778)***

TABLE VI. Different regression models fitted to PM data to predict students' final exam score. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	Model 11	Model 12	Model 13	Model 14
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Gender (female = 1, male = 0)	−0.285(0.043)***			−0.227(0.040)***	−0.009 (0.040)		−0.019 (0.038)	−0.010 (0.038)	−0.010 (0.041)	−0.019 (0.038)	0.009 (0.041)	−0.008 (0.038)	0.012 (0.041)	0.012 (0.043)
Math SAT or ACT		0.382 (0.018)***		0.375 (0.018)***		0.259 (0.017)***	0.259 (0.017)***	0.224 (0.021)***	0.259 (0.017)***	0.273 (0.019)***	0.228 (0.021)***	0.237 (0.022)***	0.274 (0.019)***	0.242 (0.022)***
Pretest CI			0.466 (0.017)***		0.465 (0.018)***	0.384 (0.017)***	0.381 (0.018)***	0.387 (0.018)***	0.365 (0.020)***	0.372 (0.019)***	0.375 (0.020)***	0.375 (0.019)***	0.354 (0.020)***	0.363 (0.021)***
Gender * math SAT or ACT								0.102 (0.035)***			0.089 (0.036)**	0.119 (0.036)***		0.104 (0.046)**
Gender * pretest CI									0.090 (0.044)**		0.060 (0.046)		0.095 (0.044)**	0.062 (0.047)
Math SAT or ACT * pretest CI										0.035 (0.020)*		0.048 (0.020)**	0.037 (0.020)*	0.050 (0.023)**
Gender * math SAT or ACT * pretest CI														−0.005 (0.049)
Constant	0.078 (0.023)***	0.000 (0.018)	0.000 (0.017)	0.062 (0.021)***	0.002 (0.020)	0.000 (0.016)	0.005 (0.020)	0.006 (0.020)	0.008 (0.020)	−0.006 (0.021)	0.007 (0.020)	−0.010 (0.021)	−0.004 (0.021)	−0.008 (0.021)
AIC	7535.7	7159	6924.3	7128.8	6926.2	6712.9	6714.6	6708.3	6712.6	6713.6	6708.6	6704.5	6711	6706.7
Observations	2669	2669	2669	2669	2669	2669	2669	2669	2669	2669	2669	2669	2669	2669
R^2	0.016	0.146	0.218	0.156	0.218	0.278	0.278	0.280	0.279	0.279	0.281	0.282	0.280	0.282
Adjusted R^2	0.016	0.145	0.217	0.155	0.217	0.277	0.277	0.279	0.278	0.278	0.279	0.280	0.279	0.280
Residual standard error	0.992 (d.o.f. = 2667)	0.924 (d.o.f. = 2667)	0.885 (d.o.f. = 2667)	0.919 (d.o.f. = 2666)	0.885 (d.o.f. = 2666)	0.850 (d.o.f. = 2666)	0.850 (d.o.f. = 2665)	0.849 (d.o.f. = 2664)	0.850 (d.o.f. = 2664)	0.850 (d.o.f. = 2664)	0.849 (d.o.f. = 2663)	0.848 (d.o.f. = 2663)	0.849 (d.o.f. = 2663)	0.848 (d.o.f. = 2661)
F statistic	43.904 (d.o.f. = 1; 2667)***	454.860 (d.o.f. = 1; 2667)***	741.816 (d.o.f. = 1; 2667)***	246.250 (d.o.f. = 2; 2666)***	370.800 (d.o.f. = 2; 2666)***	512.596 (d.o.f. = 2; 2666)***	341.713 (d.o.f. = 3; 2665)***	259.091 (d.o.f. = 4; 2664)***	257.596 (d.o.f. = 4; 2664)***	257.252 (d.o.f. = 4; 2664)***	207.666 (d.o.f. = 5; 2663)***	208.794 (d.o.f. = 5; 2663)***	206.986 (d.o.f. = 5; 2663)***	149.379 (d.o.f. = 7; 2661)***

TABLE VII. Different regression models fitted to HSEC data using additional demographic data. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	Model 1	Model 2	Model 3
	(1)	(2)	(3)
Gender (female = 1, male = 0)	−0.218 (0.072)***		−0.042 (0.072)
URM	−0.398 (0.095)***		−0.121 (0.105)
FG	−0.142 (0.155)		−0.078 (0.148)
Math SAT or ACT		0.196 (0.048)***	0.162 (0.057)***
Pretest CI		0.342 (0.048)***	0.330 (0.051)***
Constant	−0.288 (0.116)**	−0.002 (0.047)	−0.110 (0.114)
AIC	1056.4	999.9	1004.1
Observations	378	377	377
R^2	0.062	0.185	0.189
Adjusted R^2	0.055	0.180	0.178
Residual standard error	0.971 (d.o.f. = 374)	0.905 (d.o.f. = 374)	0.907 (d.o.f. = 371)
F statistic	8.287 (d.o.f. = 3; 374)***	42.360 (d.o.f. = 2; 374)***	17.243 (d.o.f. = 5; 371)***

more variables to a regression model only if that addition would improve the model fit significantly, i.e., if that addition would significantly increase the percentage of variance of dependent variable that can be explained by the model (R-squared of the model). One uses ANOVA to statistically compare the fit of multiple nested models. Models are nested if variables included in a simpler model are a subset of variables included in the more complex model(s). If the models are not nested, then one can compare the values of the AIC index of the models. The smaller the AIC of a model, the better the model fit. Variables can also produce a statistically significant improvement of the model fit without having practical educational significance. For example, an additional variable that changes the R-squared of the model from 0.29 to 0.30 has little practical significance.

Tables IV–VI capture this model evaluation process. They show different regression models for predicting final exam score using gender, math SAT or ACT score, pretest CI, and different interaction terms between these main factors. Each column represents a model. Each filled cell in a column represents a coefficient of an included variable in the model, with standard deviation of the coefficient presented in parenthesis. Below each column, AIC and the adjusted R-squared of the model along with other model statistics are reported. First, we started with single-variable regression models to predict the final exam score, e.g., predicting final exam only by gender. Then we used two-variable regression models to predict the final exam score. In the next step, we used all the three basic factors of gender, math SAT or ACT score, and pretest CI to predict final exam. Finally, to this basic additive model, we added two-way interactions as well as the three-way interactions between all the factors, and tested whether any of these additions improved the model fit. The only interaction term found to improve the model significantly was an interaction between CI prescore and math SAT or ACT score at PM, but there is no educational significance to this finding.

TABLE VIII. Different regression models fitted to HSWC 2018 data using additional demographic data. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	Model 1	Model 2	Model 3
	(1)	(2)	(3)
Gender (female = 1, male = 0)	−0.298 (0.098)***		−0.060 (0.086)
URM	−0.251 (0.111)**		0.024 (0.097)
FG	−0.453 (0.132)***		−0.111 (0.116)
Math SAT or ACT		0.340 (0.047)***	0.331 (0.050)***
Pretest CI		0.345 (0.047)***	0.336 (0.048)***
Constant	0.309 (0.077)***	−0.000 (0.041)	0.042 (0.068)
AIC	1095.2	957.3	962
Observations	394	394	394
R^2	0.078	0.347	0.349
Adjusted R^2	0.071	0.344	0.341
Residual standard error	0.964 (d.o.f. = 390)	0.810 (d.o.f. = 391)	0.812 (d.o.f. = 388)
F statistic	10.960 (d.o.f. = 3; 390)***	103.835 (d.o.f. = 2; 391)***	41.607 (d.o.f. = 5; 388)***

TABLE IX. Different regression models fitted to PM data using additional demographic data. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
	(1)	(2)	(3)	(4)	(5)	(6)
Gender (female = 1, male = 0)	−0.284 (0.043)***		−0.020 (0.038)	−0.023 (0.038)	−0.024 (0.038)	−0.025 (0.038)
URM	−0.111 (0.046)**		−0.003 (0.041)	0.0001 (0.041)	−0.0002 (0.041)	0.001 (0.041)
FG	−0.358 (0.051)***		−0.113 (0.045)**	−0.142 (0.047)***	−0.136 (0.047)***	−0.121 (0.048)**
Math SAT or ACT		0.259 (0.017)***	0.250 (0.018)***	0.278 (0.021)***	0.280 (0.022)***	0.298 (0.023)***
Pretest CI		0.384 (0.017)***	0.379 (0.018)***	0.374 (0.018)***	0.369 (0.019)***	0.353 (0.021)***
FG * math SAT or ACT				−0.089 (0.037)**	−0.096 (0.038)**	−0.115 (0.043)***
FG * pretest CI					0.036 (0.050)	0.051 (0.052)
Math SAT or ACT * pretest CI						0.046 (0.024)*
FG * math SAT or ACT * pretest CI						−0.050 (0.046)
Constant	0.166 (0.026)***	0.000 (0.016)	0.026 (0.023)	0.024 (0.023)	0.024 (0.023)	0.009 (0.024)
AIC	7476.8	6712.9	6712.1	6708.3	6709.8	6710
Observations	2669	2669	2669	2669	2669	2669
R^2	0.039	0.278	0.280	0.281	0.281	0.282
Adjusted R^2	0.038	0.277	0.278	0.280	0.279	0.280
Residual standard error	0.981 (d.o.f. = 2665)	0.850 (d.o.f. = 2666)	0.850 (d.o.f. = 2663)	0.849 (d.o.f. = 2662)	0.849 (d.o.f. = 2661)	0.849 (d.o.f. = 2659)
F statistic	36.177 (d.o.f. = 3; 2665)***	512.596 (d.o.f. = 2; 2666)***	206.677 (d.o.f. = 5; 2663)***	173.504 (d.o.f. = 6; 2662)***	148.764 (d.o.f. = 7; 2661)***	116.200 (d.o.f. = 9; 2659)***

For the HSWC 2018, HSEC, and PM data, we had URM and FG status in addition to gender. Therefore, we conducted regression analysis including these additional factors as well as gender, math SAT or ACT score, and CI prescore. In this analysis, if a factor did not have a significant main-effect contribution, we did not consider it for an interaction term. Tables VII–IX show a selection of models fitted to the data. It is clear from the regression models in Tables IV–IX that predicting final exam scores using math SAT or ACT score and CI prescores is far better than using demographics variables alone—the former explains 3–7 times more of the exam variance across the three institutions than the latter model.

APPENDIX B: STRUCTURAL EQUATION MODELING

We used structural equation modeling to test a mediation model for each institution. These models show how math SAT or ACT and CI prescore mediate the observed differences in final exam scores across demographic groups, as well as the size of the respective mediating effects for each demographic group. In the following models (Fig. 3), we first show how gender predicts both math SAT or ACT and CI prescore. We also show how those two measures of incoming preparation are correlated. Finally, we show how student final exam score is predicted by both math SAT or ACT and CI prescore. Therefore, the gender effect on the final exam score is mediated through both math SAT or ACT score and CI prescore. This model fits the data well for all institutions, as all the fit indices are within the acceptable range [root mean square error (RMSEA), acceptable range 0–0.07; comparative fit index (CFI), acceptable range above 0.95; standardized root mean square residual (SRMR), acceptable range 0–0.1], and the estimated covariances by the models were not significantly different from the actual covariances in the data, as suggested by insignificant χ^2 statistics of the models (the null hypothesis in this case is that the model is a good fit). It is notable that this model does not include any direct effect of gender on exam score, which suggests that after controlling for the effect of math SAT or ACT and CI prescore on the final exam there is no significant gender difference in final exam performance.

We added URM and FG status to the SEM model to test how measures of incoming preparation mediated the effect of these students' demographics on final exam score. Figure 4 illustrates the SEM analysis including URM and FG status as well as students' gender. For all the three institutions, the SEM model was a good fit for data, as all the fit indices were within an acceptable range, and χ^2 values of the models were insignificant.

Based on the results of these mediation models, for all three institutions, the differences in incoming preparation of underrepresented demographic groups (URM, FG, and female students) mediated the difference in their final exam scores. After controlling for the effects of math SAT or ACT and CI prescore on final exam score, there was no significant difference between the final exam score of underrepresented students and their majority peers. One exception was that after controlling for the effects of math SAT or ACT and CI prescore on exam, there remained a significant but smaller difference between final exam score of FG students and continuing generation students at PM.

For HSEC [Fig. 4(a)], math SAT or ACT and CI prescore fully mediated the effect of URM on final exam score. URM students had lower math SAT or ACT scores on average, and these scores mediated final exam scores both directly and indirectly via their effect on the CI prescore in the model. The effect of gender on final exam score was also fully mediated by CI prescores. Female students on average had lower CI prescore. CI prescore was positively

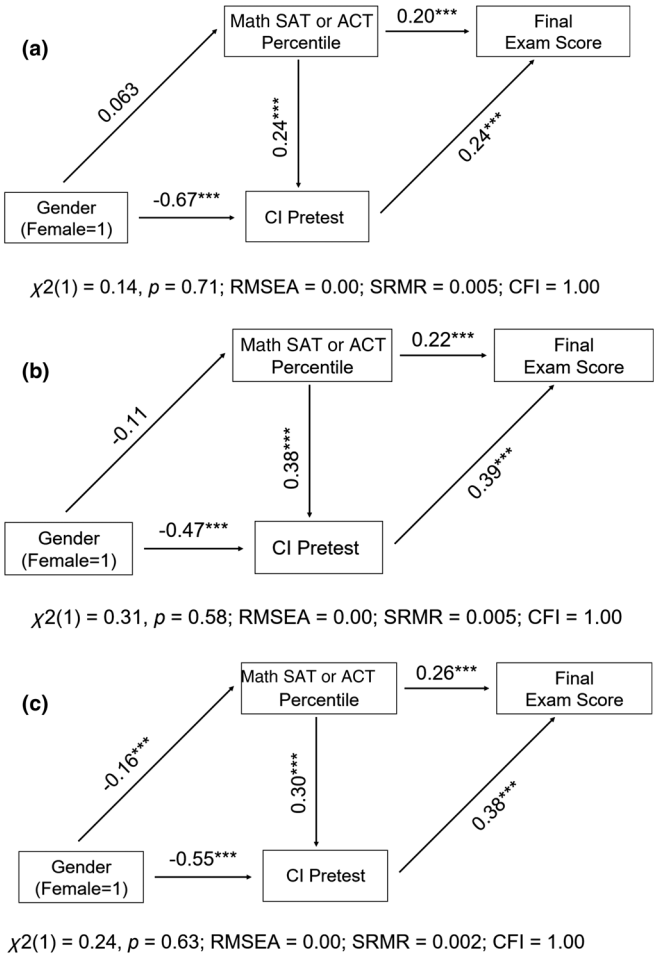

FIG. 3. The SEM models for (a) HSEC, (b) HSWC, and (c) PM data, considering gender as the only demographic variable. The arrows represent predictive relationships—e.g., gender is a predictor of math SAT or ACT score—and the numbers associated with the arrows are the effect size for the relationship. Various goodness-of-fit measures are given at the bottom of each panel. $***p < 0.001$.

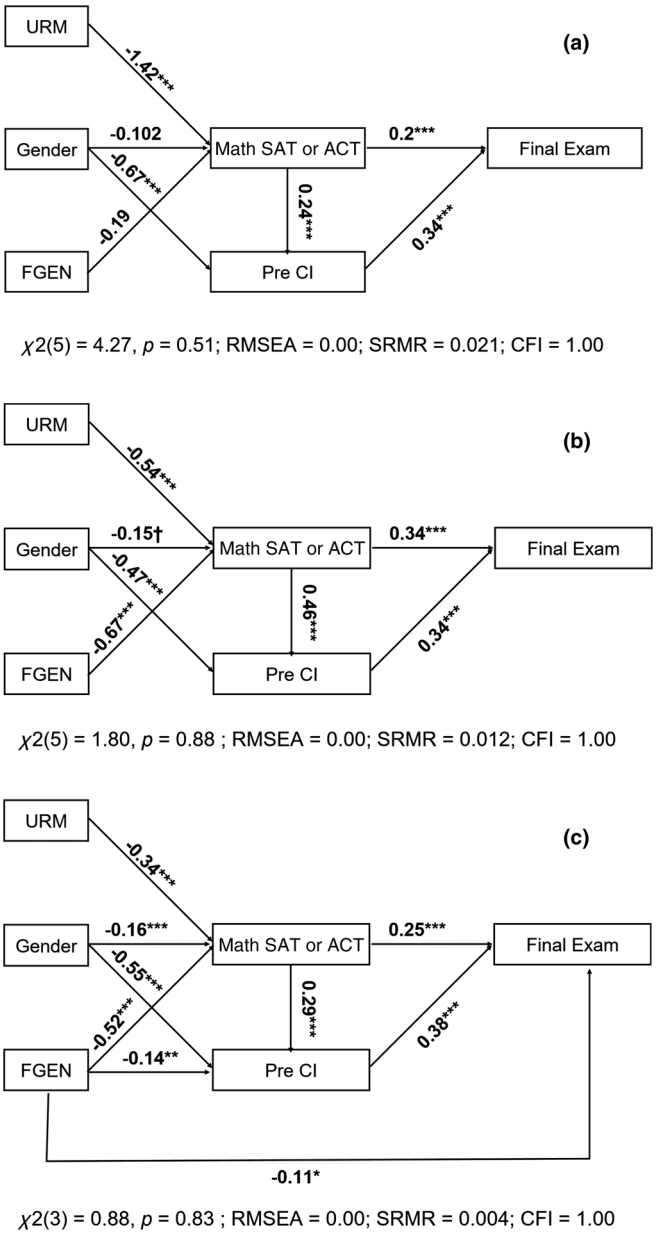

FIG. 4. The SEM models for (a) HSEC, (b) HSWC, and (c) PM data, including all demographic variables. The arrows represent predictive relationships—e.g., gender is a predictor of math SAT or ACT score—and the numbers associated with the arrows are the effect size for the relationship. Various goodness-of-fit measures are given at the bottom of each panel. *** $p < 0.001$.

correlated with final exam score. For FG students there was no significant gap in performance, and therefore, no mediation effect through measures of incoming preparation.

For HSWC [Fig. 4(b)], math SAT or ACT and CI prescore fully mediated the effect of URM on final exam score. URM and FG students had on average lower math SAT or ACT scores, and these scores mediated final exam scores both directly and indirectly via their effect on the CI prescores in the model. The effect of gender on final exam score was also fully mediated by incoming preparation, both directly through lower math SAT or ACT scores and CI prescores, as well as indirectly by math SAT score via CI prescore. Female students on average had both lower math SAT or ACT and lower CI prescore. Both of these scores were positively correlated with final exam score. Furthermore, math SAT or ACT score was positively correlated with CI prescore. Therefore, lower math SAT or ACT scores of female students not only directly mediated the effect of gender on final exam performance, but also indirectly mediated through correlation with CI prescore and the CI prescore being correlated with final exam scores.

For PM [Fig. 4(c)], math SAT or ACT and CI prescore fully mediated the effect of URM and gender on final exam score. URM had on average lower math SAT or ACT scores, and these scores mediated final exam scores both directly and indirectly via their effect on the CI prescore in the model. Both of these scores were positively correlated with final exam scores. Furthermore, math SAT or ACT score was positively correlated with CI prescore. Therefore, lower math SAT or ACT scores of female students not only directly mediated the effect of gender on final exam performance, but also indirectly mediated through correlation with CI prescores and these CI scores being correlated with final exam scores. The effect of gender on final exam score was also fully mediated by incoming preparation, both directly through lower math SAT or ACT score and CI prescore as well as the indirect effect of math SAT score via CI prescore. Female students on average had both lower math SAT or ACT and lower CI prescores. Both of these scores were positively correlated with final exam scores. Furthermore, math SAT or ACT score was positively correlated with CI prescore. Therefore, lower math SAT or ACT scores of female students not only directly mediated the effect of gender on final exam performance, but also indirectly mediated through correlation with CI prescores and these CI scores being correlated with final exam scores. The effect of FG on final exam was also partially mediated by incoming preparation, both directly through lower math SAT or ACT scores and CI prescores, as well as the indirect effect of math SAT scores via CI prescores. This was partial mediation, as after controlling for the effect of math SAT or ACT and CI prescore on final exam, there existed a significant but smaller FG performance gap in final exam.

APPENDIX C: FAILURE ANALYSIS AND MULTIPLE REGRESSION ANALYSIS

The connection between the failure analysis and multiple regression analysis is subtle, but general. For the case of input variables that are reasonably close to normal distributions, this analysis can be reduced to a dependence purely on the value of R-squared in the regression model.

Our multiple linear regression estimates the linear relationship between final exam score and a linear

combination of math SAT or ACT and CI prescore:

$$z_f = b_{\mathrm{Math}} z_{\mathrm{Math}} + b_{\mathrm{CI}} z_{\mathrm{CI}} + \epsilon, \tag{C1}$$

where b_i are the regression coefficients in Table III, and ϵ is the residual error. All variables have been converted to z scores. Assuming these z scores are all normally distributed, we can define a single composite z score for preparation,

$$z_p = \frac{b_{\mathrm{CI}} z_{\mathrm{CI}} + b_{\mathrm{Math}} z_{\mathrm{Math}}}{\sqrt{b_{\mathrm{CI}}^2 + b_{\mathrm{Math}}^2}}, \tag{C2}$$

that is also normally distributed, and carries a fraction R^2 of the variance in z_f: $z_f = \sqrt{R^2} z_p + \epsilon$. Therefore, $\epsilon = z_f - \sqrt{R^2} z_p$ is normally distributed with variance $1 - R^2$, and the joint probability distribution of z_p and z_f is

$$P(z_p, z_f) = \frac{1}{2\pi\sqrt{1-R^2}} \exp\left(-\frac{z_p^2}{2} - \frac{(z_f - \sqrt{R^2} z_p)^2}{2(1-R^2)} \right). \tag{C3}$$

The bottom quartile in a normal distribution consists of all values that are more than $Q\sigma$ below the mean, where $Q = 0.674$. Thus, the failure rate (bottom quartile of quartile of z_f) for students in the bottom quartile of the preparation scores z_p is

$$F_L = C_L \int_{-\infty}^{-Q} dz_p \int_{-\infty}^{-Q} dz_f P(z_p, z_f), \tag{C4}$$

where the normalization coefficient C_L is

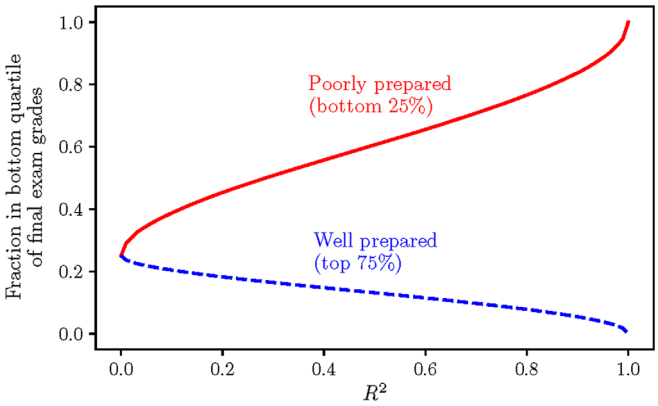

FIG. 5. Fraction of students in the bottom quartile of exam scores as a function of R-squared for the prior preparation model in Table III.

$$C_L^{-1} = \int_{-\infty}^{-Q} dz_p \int_{-\infty}^{\infty} dz_f P(z_p, z_f) = \frac{1}{4}.$$

Since the total "failure" rate (below the 25th percentile on the final exam) for the whole class is 0.25 by definition, the failure rate for students in the top 75% of incoming preparation is $F_U = (1 - F_L)/3$. This result is plotted in Fig. 5.

The above analysis assumes normally distributed measures of incoming preparation, which is not true of most such data, including our example data for HSWC 2018. However, the logistic regression analysis, which does not have that limitation on the distribution of the incoming preparation data, gives nearly identical results, indicating that this relationship between R-squared and the probability of failure is quite robust to violations of this assumption.

[1] C. H. Crouch and E. Mazur, Peer instruction: Ten years of experience and results, Am. J. Phys. **69**, 970 (2001).

[2] K. Cummings, J. Marx, R. Thornton, and D. Kuhl, Evaluating innovation in studio physics, Am. J. Phys. **67**, S38 (1999).

[3] R. Hake, Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses, Am. J. Phys. **66**, 64 (1998).

[4] S. Freeman, S. L. Eddy, M. McDonough, M. K. Smith, N. Okoroafor, H. Jordt, and M. P. Wenderoth, Active learning increases student performance in science, engineering, and mathematics, Proc. Natl. Acad. Sci. U.S.A. **111**, 8410 (2014).

[5] M. Lorenzo, C. H. Crouch, and E. Mazur, Reducing the gender gap in the physics classroom, Am. J. Phys. **74**, 118 (2006).

[6] R. J. Beichner, J. M. Saul, D. S. Abbott, J. J. Morse, D. Deardorff, R. J. Allain, and J. S. Risely, The student-centered activities for large enrollment undergraduate programs (SCALE-UP) project, Rev. Phys. Educ. Res. **1**, 2 (2007), http://www.per-central.org/document/ServeFile.cfm?ID=4517.

[7] S. Freeman, E. O'Connor, J. W. Parks, M. Cunningham, D. Hurley, D. Haak, and M. P. Wenderoth, Prescribed active learning increases performance in introductory biology, CBE Life Sci. Educ. **6**, 132 (2007).

[8] D. C. Haak, J. HilleRisLambers, E. Pitre, and S. Freeman, Increased structure and active learning reduce the achievement gap in introductory biology, Science **332**, 1213 (2011).

[9] C. J. Ballen, C. Wieman, S. Salehi, J. B. Searle, and K. R. Zamudio, Enhancing diversity in undergraduate science:

Self-efficacy drives performance gains with active learning, CBE Life Sci. Educ. **16**, ar56 (2017).
[10] R. Fullilove and P. Treisman, Mathematics achievement among African American undergraduates at the University of California, Berkeley: An evaluation of the Mathematics Workshop program, J. Negro Educ. **59**, 463 (1990).
[11] R. H. Tai and P. M. Sadler, Gender differences in introductory undergraduate physics performance: University physics versus college physics in the USA, Int. J. Sci. Educ. **23**, 1017 (2001).
[12] T. G. Greene, C. N. Marti, and K. McClennney, The effort-outcome gap: Differences for African American and Hispanic community college students in student engagement and academic achievement, J. Higher Educ. **79**, 513 (2008).
[13] 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. ST Phys. Educ. Res. **9**, 020121 (2013).
[14] S. L. Eddy and K. Hogan, Getting under the hood: How and for whom does increasing course structure work?, CBE Life Sci. Educ. **13**, 453 (2014).
[15] S. L. Eddy, S. E. Brownell, and M. P. Wenderoth, Gender gaps in achievement and participation in multiple introductory biology classrooms, CBE Life Sci. Educ. **13**, 361 (2014).
[16] *Women, minorities, and persons with disabilities in science and engineering* (National Science Foundation, Arlington, VA, 2015), https://ncses.nsf.gov/pubs/nsf19304/digest/field-of-degree-women.
[17] S. L. Eddy and S. E. Brownell, Beneath the numbers: A review of gender disparities in undergraduate education across science, technology, engineering, and math disciplines, Phys. Rev. Phys. Educ. Res. **12**, 020106 (2016).
[18] J. B. Hinnant, M. O'Brien, and S. R. Ghazarian, The longitudinal relations of teacher expectations to achievement in the early school years, J. Educ. Psychol. **101**, 662 (2009).
[19] C. M. Steele and J. Aronson, Contending with a stereotype: African-American intellectual test performance and stereotype threat, J. Pers. Soc. Psychol. **69**, 797 (1995).
[20] L. E. Kost, S. J. Pollock, and N. D. Finkelstein, Unpacking gender differences in students' perceived experiences in introductory physics, AIP Conf. Proc. **1179**, 177 (2009).
[21] J. Watkins, Examining issues of underrepresented minority students in introductory physics. Ph.D. thesis, Harvard University, 2010.
[22] E. Brewe, V. Sawtelle, L. H. Kramer, G. E. O'Brien, I. Rodriguez, and P. Pamelá, Toward equity through participation in Modeling Instruction in introductory university physics, Phys. Rev. ST Phys. Educ. Res. **6**, 010106 (2010).
[23] B. Wilcox and H. Lewandowski, Research-based assessment of students' beliefs about experimental physics: When is gender a factor?, Phys. Rev. Phys. Educ. Res. **12**, 020130 (2016).
[24] 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).
[25] L. E. Kost, S. J. Pollock, and N. D. Finkelstein, Characterizing the gender gap in introductory physics, Phys. Rev. ST Phys. Educ. Res. **5**, 010101 (2009).
[26] L. E. Kost-Smith, S. J. Pollock, and N. D. Finkelstein, Gender disparities in second-semester college physics: The incremental effects of a "smog of bias", Phys. Rev. ST Phys. Educ. Res. **6**, 020112 (2010).
[27] 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).
[28] 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).
[29] 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).
[30] D. Hestenes, M. Wells, and G. Swackhamer, Force Concept Inventory, Phys. Teach. **30**, 141 (1992).
[31] L. Staffaroni, Historical SAT percentiles: New SAT 2016, 2017, and 2018, https://blog.prepscholar.com/historical-percentiles-new-sat.
[32] H. Akaike, Likelihood of a model and information criteria, J. Econometrics **16**, 3 (1981).
[33] Y. Rosseel, Lavaan: An R package for structural equation modeling and more, version 0.5–12 (BETA), J. Stat. Softw. **48**, 1 (2012).
[34] R Core Team R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org/.
[35] D. Gefen, D. W. Straub, and M. Boudreau, Structural equation modeling and regression: Guidelines for research practice, Commun. Assoc. Inf. Syst. **4**, 7 (2000).
[36] W. K. Adams, K. K. Perkins, M. Dubson, N. D. Finkelstein, and C. E. Wieman, AIP Conf. Proc. **790**, 45 (2005).
[37] M. Scott, T. Stelzer, and G. Gladding, Evaluating multiple-choice exams in large introductory physics courses, Phys. Rev. ST Phys. Educ. Res. **2**, 020102 (2006).
[38] J. Day, J. B. Stang, N. G. Holmes, D. Kumar, and D. A. Bonn, Gender gaps and gendered action in a first-year physics laboratory, Phys. Rev. Phys. Educ. Res. **12**, 020104 (2016).
[39] S. D. Willoughby and A. Metz, Exploring gender differences with different gain calculations in astronomy and biology, Am. J. Phys. **77**, 651 (2009).
[40] L. C. McDermott and P. S. Shaffer, *Tutorials in Introductory Physics* (Upper Saddle River, Prentice-Hall, NJ, 2002).