Grades, grade component weighting, and demographic disparities in introductory physics

Amber B. Simmons and Andrew F. Heckler
Department of Physics, Ohio State University, Columbus, Ohio 43210, USA

(Received 15 May 2020; accepted 11 August 2020; published 21 October 2020)

Two set of studies were conducted to better understand grades and grading practices in physics courses, and how these might influence demographic representational disparities in physics. The first study investigates the relationships between grades and the student-level factors of standardized test scores, (binary) gender, underrepresented minority (URM) status, first generation (FG) status, citizenship status, and age of over 20 000 students enrolled in algebra-based and calculus-based introductory physics courses. Consistent with other studies, we find differences in mean grades for all of these factors, except for gender, and when standardized test scores are included in a regression model predicting grades, the demographic differences in grades decreases, though typically remain nonzero. We also find gender by test score and URM by test score interactions when predicting grades. The second study examines grade component scores, and replicates the finding that compared to men, women achieve higher scores on nonexam components and lower scores on exam components. We also find that the gap in score between URM and FG students and their counterparts is less for non-exam components than for exam components. Because of these differentials in components, we compared different models of grade components weighting and find that women and URM students differentially benefit from stronger weighting of nonexam components. While the benefit to grades is relatively small, the relative shift in percentages of grade rates of A, D, and F can have dramatic differential shifts. We also find that while exam components are moderately strongly correlated with standardized tests scores, nonexam components are not. These results suggest that grade component weighting is inevitably tied to issues of demographic equity, in the sense that altering the weights may change demographic disparities in grades and change the dependency of grades on standardized test scores. We conclude with a call for more attention to grading practices and what is rewarded in introductory physics courses.

DOI: 10.1103/PhysRevPhysEducRes.16.020125

I. INTRODUCTION

It is readily noticed and well documented that women, black or African Americans, Hispanic or Latinos, and American Indian or Alaskan Natives are substantially underrepresented—often by more than a factor of 2—in university enrollment and degrees received in the physical sciences, especially in physics [1–3]. What is less straightforward is why. The vast number of papers and projects that have been dedicated to understanding why this underrepresentation occurs or how to address it suggests that there is a complex array of causes and contexts. The focus of this paper is on the topic of grades and grade components in introductory physics courses in order to gain more insight as to whether grades and grading practices may indicate or compound issues of demographic biases or may play a role in demographic disparities in participation. It is important to note that we are investigating grades not because we suppose they are a measure of learning, but because they are valued by students, instructors, and programs, and they play an important role in continued participation in programs. Grades are concrete outcomes with real-world consequences for students, and, for better or worse, grades provide feedback for and inform decisions by students, instructors, and programs. We begin with a discussion of relevant prior work.

Historically, it had been somewhat common and compelling to put forth the argument that the demographic disparities in science, technology, engineering, and mathematics (STEM) participation are due to differences in preparation or prior achievement (e.g., see a brief critical review of this perspective in Ref. [4]). However, several researchers have shown that, at best, demographic differences in prior achievement can only account for a portion of the disparity in participation. For example, controlling for academic preparation, women and minorities still disproportionally enter and leave STEM physical science and engineering major programs [4–10]. Further,

Published by the American Physical Society under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

even for the studies that find some lower relative grades or higher failure rates for women and minorities compared to white or Asian males for a given physics course, the differences are relatively small such that they are unlikely to be the major contributor to that massive difference in participation (see, for example, Ref. [3]). Therefore, there must be other more important factors besides preparation causing disparities in participation.

In fact, in contrast to considering deficits in science achievement as the cause for disparities in participation, a number of researchers have pointed out that, compared to men, women commonly have high levels of achievement in verbal skills and the humanities. Thus it has been proposed that, for example, women with high levels of math and science achievement also have, compared to men, relatively more viable options for careers in other fields, thus are more likely to leave STEM, especially areas that are male dominated and perceived as relatively unwelcoming to women (see, for example, Refs. [4,6–8,11]).

Grades are also a natural and compelling factor to consider as related to, and perhaps responsible for, demographic disparities in STEM participation. For this purpose, a number of researchers have investigated whether there are significant differences in university-level physics course grades among various demographic groups. When comparing grade performance by gender, the results are mixed with a relatively slight overall trend indicating that, while there is some variation favoring either side, men score roughly 0.1 grade points higher than women on the four-point grade scale averaged over many classes and institutions [12]. The variation may be due to unidentified contextual factors; for example, Lauer *et al.* [13] found no grade differences in physics. Hazari *et al.* [9] found that men had a similar level of higher grades even when controlling for numerous prior factors such as standardized test scores. Kost, Pollock, and Finkelstein [14–15] also found a similar small grade difference but propose that this difference—along with other small differences in attitudes and beliefs—accumulate over time to significantly influence participation. When comparing grades by underrepresented minority status, the differences are somewhat larger and thus more potentially impactful, averaging around 0.2–0.4 grade points lower for minorities [9,16–17], controlling for various factors such as standardized tests scores. In sum, there are differences in grades on average between women and men and minorities and nonminorities in directions consistent with disparities in participation, but the differences are relatively small, are reduced when other factors such as ACT scores are taken into account, and are thus likely to account for only a portion of the disparities in participation.

What may be more important to consider is the extent to which students use grades as feedback for deciding whether to stay in a particular STEM major program [4,6,8,10, 18–20]. There is evidence that all students use grades to revise their own beliefs about their ability in a field (e.g., Ref. [21]), but the important point here is whether there are differential sensitivities between demographic groups in which grades plays a role in driving disparities in STEM retention. For example, in modeling a large national dataset of approximately 9000 students, Astorn-Figari and Speer [18] provide some evidence that students with low grades switch majors more, and all students tend to switch to majors with demographics similar to their own. Further, they found that the gender disparity in switching out of physical science and engineering is not due to gender differences in grades or measures of ability, rather women tend to switch out of these fields to go into those that are less male-dominated and where they have higher grades. Kugler *et al.* [8] found somewhat similar results with a sample of ∼9000 students from a large private university: both women and men are equally more likely to switch majors if they have low grades, and this is even true regardless of gender composition of the major *except* in the case of male-dominated STEM majors. In that case women are more likely than men to leave such majors. They hypothesized that the context of external stereotype threat in male-dominated STEM majors is a critical additional cue signaling "lack of fit" to women and causes the disparity. Using more direct analysis of persistence in a major, Ost [10] samples ∼15 000 students from a "large elite research university" and found that students are "pulled away" from STEM by their high grades in non-STEM and "pushed away" by their low grades in STEM, and among physical science majors, not only do females tend to have higher grades in non-STEM courses, but they are also found to be more responsive to grades than males, consistent with theories of stereotype vulnerability. In studying over 8000 students who took General Chemistry I, Witherspoon *et al.* [6] do not find such differences in retention between genders for students receiving lower grades, but they do find that competency beliefs moderate a gendered difference in enrollment in 2nd year chemistry for students receiving an A or B. They hypothesize that because of the presence of stereotype threat in science courses, women who receive high grades also perceive the relatively high effort to achieve such grades (compared to non-STEM courses), and this may be attributed by them as evidence of lack of ability. Ahn *et al.* [19] model data from 16 000 students from the University of Kentucky and also find, as expected, that grades play an important role in major choice, but that women value grades more than men (see also for example Rask and Tiefenthaler [22]). Since they find that STEM fields award significantly lower grades than non-STEM fields, their models suggest that parity in average grades among departments in a university would significantly increase gender participation in STEM majors.

Since grades appear to play an important but nuanced role in demographic disparities in participation, understanding differences in performance in the components of the course that make up the grade may provide important

insights into the mechanisms driving disparities. In short, there is substantial evidence that differences in performance on specific grade components exist between demographic groups. Kost *et al.* [14–15] examined exam, homework, and participation scores over numerous semesters and consistently found that women scored higher than men on homework and participation scores and lower on exams. These differences offset each other, resulting in small or negligible differences in course grade. Kortemeyer [23] found results consistent with this and found small differences in how men and women interact with the online homework assignments. For example, he found that women more frequently report using the multiple attempt feature of the assignment to explore their own thinking about their own problem-solving approaches without worrying about the score while men more frequently report that they use the multiple attempts for guessing.

In a more comprehensive study, Salehi *et al.* [24] investigated the grade component performance of ∼6000 students in STEM courses including biology, physical science, and engineering courses. They reported a significant and persistent "gender penalty" for women in exam performance in first- or second-year physical science or engineering courses. That is, women tended to score lower on exams than men. In contrast, for lab scores, and in-class scores such as participation and groupwork, women performed the same (or better) than men. When ACT scores were included in the model, these differences remained. Further, ACT score was a much weaker predictor of nonexam and lab scores, with beta estimates for ACT scores about one-third as large for non-exams scores than for exam scores. They also found that when ACT scores were taken into account, underrepresented minorities (URM) tended to score about the same on exams as non-URM students, but lower on nonexam components and labs, while first-generation (FG) students on average scored the same as their counterparts. Salehi *et al.* also provided evidence that the moderately strong association of exam scores with ACT scores was mediated (though somewhat weakly) by test anxiety for women (and not men), and this may partially explain the disparity in exam performance.

In sum, grades themselves appear to play an important role in students' perception of ability and their subsequent choices for a major, and these perceptions and choices can depend on the demographic group and the perceived climate of the major for that group. Further, mean performance on a given grade component varies by demographic group, and this variation varies by grade component. Exam grade components, which typically comprise the majority of the final grade, are fairly strongly correlated with ACT scores, while other grade components are not.

This variation of grade component performance with demographic group raises important questions about how various grade components should be weighted to maximize fairness and represent what is valued by the instructors, the major program, and the students. For introductory physics courses at Ohio State University, tests (exams and quizzes) comprise 70% of the grade, and informal conversations with colleagues at other institutions indicate that this majority weighting of tests is common. As discussed, those that score highest on exams are non-URM males with high ACT scores. To what extent do grade disparities change when the weighting is changed?

In this paper we will investigate grades and grade components of students in algebra-based and calculus-based introductory physics courses at Ohio State University, a large public research university. Because prior research such as Matz *et al.* [12] report some variation in demographic disparities, we will first characterize grade outcomes in our local context, including demographic differences in grades and grade components. Then we will present several models that explore hypothetical outcomes of a slightly different grade weighting scheme that reduces the weights of tests—and, consequently, ACT scores—on grades. In short, this study seeks to answer the following research questions:

(1) What are the mean ACT math scores, ages, and grades by gender, URM status, first-generation status, and international student status?
(2) To what extent are there demographic differences in grade performance controlling for ACT scores, age, and lecture section? Are there significant interactions of demographic groups with ACT score when comparing grades?
(3) To what extent are there demographic differences in performance on grade components, both controlling and not controlling for ACT scores?
(4) If the test components are weighted less than the current weight (70%), how does this change the answers to research questions 1 and 2?

Overall, this paper will contribute to the field in two ways. First, it provides a novel exploration of the effect of changing the weighting of grade components on grade outcomes for several important demographic groups. This will lead us to a commentary on ACT scores, grades, and grading practices. Second, we confirm and extend results of previous studies mentioned above on demographic differences in grade outcomes. Such careful reporting of results is critical in order to increase confidence in the generalizability of conclusions, especially when there is a surprising relative paucity of formal results and studies in this area, and for the data that exists there is significant variation. We also describe a wholistic picture of our local context, allowing for a richer, more specified setting of the results. Finally, note that this study is the first of two parts. In the first part reported here, we investigate student-level effects, and part two, to be presented in a separate paper, will model and discuss unexplained variance in grades due to semester, section, and instructor-level effects.

II. METHODS

A. Data

The data used in this study were collected from the university registrar records on students who were enrolled between Autumn 2012 and Spring 2018 in either the first semester of the algebra-based ($N = 9017$) or calculus-based ($N = 11\,256$) introductory physics two-semester sequence at Ohio State University. Both courses were large, traditional lecture-style courses with a typical enrollment of approximately 200 students in each lecture section. Lectures were 2–3 times per week and each lecture section was divided into smaller (~24 students) recitation meetings 1–2 times per week, where quizzes were administered and students reviewed homework and other example problems, as well as a lab class once per week. The algebra-based course consists primarily of pre-health students, and the calculus-based course is largely pre-engineering students, with a small fraction of science and math majors. Thirteen different instructors taught the 52 algebra-based lecture sections included in the 12 semesters in this dataset. Nineteen instructors taught the 60 calculus-based lecture sections offered between Autumn 2012 and Spring 2018.

In addition, grade component scores were collected for both courses from Autumn 2016 through Spring 2018 ($N = 6587$) for both the algebra-based and calculus-based introductory physics courses. The grades are composed of nonexam components (online homework, online essential skills assignments [25], labs) and test components (in-class formal quizzes, midterm exams, and final exam).

In the physics department at OSU during the time these data were collected, a grade curving system was used. In the algebra-based course, the quiz total, the first midterm, the second midterm, and the final exam each were shifted so that the median score of each was a 77% if the median was under 77%. If the median was above 77%, no curve was awarded. In the calculus-based course, each instructor determined a grade scheme for each lecture section by assigning grade cutoffs that would result in a $C+$ to $B-$ average for the course.

Frequencies for gender, URM status, FG status, and citizenship status along with ACT math score and course outcomes are reported by demographic information and course in Table I. Note the URMs were defined here as black or African Americans, Hispanic or Latinos, and American Indian or Alaskan Natives as reported in the university student database. In the algebra-based course, women represent 57% of the population. URM students comprise 10% of the students, and 19% of students are first-generation college students. In the calculus-based course, female students make up only 23% of the population, URM students represent 9% of the population, and 17% of this population are first-generation students. Since registrar records only categorize gender as binary, we are unable to provide statistics on nonbinary gender populations.

Note that in this paper, the student's highest math ACT score was used. For students that reported SAT scores, we used the ACT and College Board's ACT/SAT concordance tables to determine the ACT equivalent to recorded SAT scores. About 9% of the students in this sample do not have either an ACT or SAT Math score reported. As seen in Table I, these students on average receive substantially lower course grades, a higher DFW rate, a lower A rate, and lower retention as compared to the rest of the population. Although there are several reasons a student may not have an ACT or SAT score on record, the higher mean age of this population and our informal discussions with OSU enrollment staff indicate that the overwhelming majority of these students are likely transfer students.

First-generation status is considered in this study as growing amounts of research show differences in motivation and behavior in STEM between first-generation and non-first-generation students. For example, one study of engineering students shows differences in identity, performance and competence beliefs, interests, and family support for science [26]. Another study shows particular importance of early college performance in retention of first-generation students [27].

We analyze data based on foreign or domestic status (i.e., citizenship) of the students as well. The principle reason for doing so is the considerable increase in foreign students in our institution in the last decade, and they comprise a sizable proportion of our students (~10%). Since there is very little reported on the performance of foreign students, we were interested to see if there are any important differences that may warrant attention.

Finally, the age of the student is also considered in the following analyses given the known retention issues for older students in STEM fields. For example, students who begin their college education at age 19 or younger are more likely to complete a bachelor's education in STEM than students who entered at 20 or older. Additionally, when students left a STEM field, a higher percentage of younger students changed to a non-STEM field while a higher percentage of older students left university without any degree [28].

B. Statistical analysis

Because of the nested nature of our student data (students clustered within lecture sections), multilevel modeling was used to determine demographic differences in performance outcomes controlling for ACT math, age, and the lecture section as given by

$$\begin{aligned}\mathrm{Grade}_{ij} = {} & \gamma_{00} + u_{0j} + \gamma_{10}(\mathrm{CWC\,ACT}_{ij}) \\ & + \gamma_{20}(\mathrm{CWC\,Age}_{ij}) + \gamma_{30}(\mathrm{Gender}_{ij}) \\ & + \gamma_{40}(\mathrm{URM}_{ij}) + \gamma_{50}(\mathrm{FG}_{ij}) \\ & + \gamma_{60}(\mathrm{Citizen}_{ij}) + r_{ij}, \end{aligned} \tag{1}$$

TABLE I. Descriptive performance statistics by demographics. Columns with an * represent the outcomes of students' first attempt at each course if they repeated that course. ACT math score, grades, % DFW, %A, and % retention were t tested by demographic. Bold numbers indicate $p < 0.05$ difference between the demographic groupings.

					ACT (or SAT conversion) math score														
					Age									Grade*					
		Population	Count	% of Full population	*M*	*SD*	% No Score Reported	<27	27–29	30–32	>32	*M*	*SD*	*M*	*SD*	% DFW*	% A*	% Retention
First semester algebra-based physics		Full population	9017		20.6	2.0	9.2	32.8%	30.8%	20.7%	15.7%	28.1	3.9	2.81	1.1	16.4	34.4	77.1
	Gender	Male	3855	42.8	**20.8**	2.2	**10.4**	**29.5%**	30.0%	**22.1%**	**18.4%**	**28.5**	4.0	2.80	1.2	**18.7**	**36.0**	79.1
		Female	5150	57.1	**20.5**	1.8	**8.2**	**35.2%**	31.3%	**19.7%**	**13.8%**	**27.9**	3.9	2.82	1.1	**14.6**	**33.2**	76.7
	URM status	URM	950	10.5	**20.8**	2.5	**10.5**	**57.4%**	**26.2%**	**10.8%**	**5.5%**	**25.7**	4.1	**2.36**	1.2	**28.5**	**19.0**	**67.6**
		nonURM	7604	84.3	**20.5**	1.9	**7.7**	**30.6%**	**31.8%**	**21.3%**	**16.3%**	**28.3**	3.8	**2.85**	1.1	**15.2**	**35.3**	**79.2**
	First generation status	FirstGen	1722	19.1	**20.4**	1.5	**6.4**	**47.6%**	29.5%	**14.3%**	**8.5%**	**26.7**	4.0	**2.53**	1.1	**22.5**	**23.3**	76.2
		NonFirstGen	7294	80.9	**20.6**	2.1	**9.8**	**29.2%**	31.1%	**22.2%**	**17.5%**	**28.5**	3.9	**2.88**	1.2	**14.9**	**37.0**	78.0
	Citizen status	Foreign	1286	14.3	**20.4**	1.8	**23.1**	**9.2%**	**9.9%**	**32.0%**	**48.9%**	**31.5**	3.7	**3.22**	1.0	**8.3**	**53.8**	**63.7**
		Domestic	7728	85.7	**20.6**	2.0	**6.8**	**36.0%**	**33.6%**	**19.1%**	**11.2%**	**27.7**	3.8	**2.74**	1.1	**17.7**	**31.1**	**78.5**
	No standardized test score reported		826	9.2	23.4	4.3								2.38	1.4	32.8	29.7	52.9
First semester calculus-based physics		Full population	11256		19.6	2.1	8.5	15.8%	31.0%	28.2%	25.1%	29.8	3.4	2.57	1.1	20.7	21.8	69.6
	Gender	Male	8581	76.2	**19.6**	2.3	**9.3**	**15.3%**	**29.7%**	**28.7%**	**26.3%**	**29.9**	3.5	2.56	1.1	21.0	21.8	69.6
		Female	2655	23.6	**19.3**	1.5	**5.8**	**17.4%**	**34.8%**	**26.6%**	**21.2%**	**29.5**	3.3	2.59	1.0	19.6	21.7	70.9
	URM status	URM	985	8.8	**19.7**	2.4	**10.9**	**35.3%**	33.1%	**23.0%**	**8.5%**	**27.6**	3.6	**2.11**	1.1	**35.4**	**10.6**	**61.7**
		nonURM	9701	86.2	**19.5**	2.1	**7.3**	**14.1%**	31.0%	**28.6%**	**26.4%**	**30.0**	3.3	**2.60**	1.1	**19.3**	**22.5**	**70.8**
	First generation status	FirstGen	1940	17.2	19.5	2.0	**6.9**	**25.1%**	**35.5%**	**22.3%**	**17.0%**	**28.7**	3.6	**2.32**	1.1	**27.2**	**15.7**	**62.8**
		NonFirstGen	9315	82.8	19.6	2.1	**8.9**	**13.8%**	**30.0%**	**29.4%**	**26.8%**	**30.0**	3.4	**2.62**	1.1	**19.4**	**23.1**	**71.4**
	Citizen status	Foreign	1428	12.7	**19.8**	1.9	**21.3**	**6.7%**	**10.0%**	**25.4%**	**57.9%**	**32.2**	3.3	**2.96**	1.1	**13.2**	**40.5**	73.1
		Domestic	9827	87.3	**19.5**	2.1	**6.7**	**16.9%**	**33.5%**	**28.5%**	**21.1%**	**29.5**	3.3	**2.51**	1.1	**21.8**	**19.0**	69.4
	No standardized test score reported		961	8.5	23.4	4.6								2.25	1.3	38.3	22.6	55.7

where subscript i refers to an individual student, and subscript j refers to their lecture section. Grade_{ij} refers to an individual student's grade, γ_{00} is the overall intercept or grand mean when all predictors are zero, and u_{0j} accounts for the random effect of lecture section: it is the group-level random error of the mean of group j from the grand mean, and this can account for variations by section such as time of day, lecturer, scheduling priority given to different groups, cohorts of students in programs, etc. The r_{ij} term indicates the student-level random error that is not explained by the model. Each γ term indicates the regression coefficient between its predictor and the outcome. ACT math score and age were mean-centered within cluster (CWC). Thus, CWC ACT_{ij} describes how many points a given student earned above or below the mean score within that student's course section. For each categorical demographic variable, the reference group was chosen as the majority value; for gender, male students are the reference group. For underrepresented minority status (URM), nonminority students are the reference group. For first-generation status (FG), non-first-generation students are the reference group, and for citizen status (Citizen), domestic students are the reference group. Both random intercept models and random slope models were tested, but there were no appreciable differences between the estimates in the two models as measured by the Akaike information criterion (AIC), so random intercept models were used. In this paper we have chosen to model the effects of lecture section, which can include factors such as mean ACT, mean age, time of day, and instructor, as a random effect; we are not studying any potential systematic or casual effects of such factors in this paper.

In addition to modeling the grade as an outcome, we will also model binary outcomes such as the probability of receiving a DFW (grade of D, fail, or withdraw), the probability of receiving an A in the course, or "retention" to the second course in the introductory sequence. Retention is calculated only for those students majoring in programs (while enrolled in the course) requiring the second physics course in the sequence and is defined as enrolling in the second course within one year from the last time the student took the first course (note that some students, ∼5% for the algebra-based students and ∼10% for the calculus-based students, took the first course more than once). For example, if the last time the student took the first course was in Autumn of 2015, then if they enrolled in the second course in Spring 2016, Summer 2016 or Autumn 2016, they were considered retained.

These outcomes allow us to gain more insight beyond mean scores and into information about the grade distributions. Such grades are also important outcomes for students, who need to either pass the course or receive a high grade in the course in order to enter into subsequent competitive programs such as medical school or an engineering major. For these binary outcomes, the following multilevel (i.e., levels of clustering) logistic regression model was used:

$$\begin{aligned}\text{logit}(p(\text{DFW}_{ij})) = {} & \gamma_{00} + u_{0j} + \gamma_{10}(\text{CWC ACT}_{ij}) \\ & + \gamma_{20}(\text{CWC Age}_{ij}) + \gamma_{30}(\text{Gender}_{ij}) \\ & + \gamma_{40}(\text{URM}_{ij}) + \gamma_{50}(\text{FG}_{ij}) \\ & + \gamma_{60}(\text{Citizen}_{ij}),\end{aligned} \tag{2}$$

where given a probability p of an outcome, $\text{logit}(p) = \log[p/(1-p)]$, or

$$p = \text{logit}^{-1}(p) = \frac{1}{1+e^{-p}}. \tag{3}$$

We also examined interactions between demographic group and ACT score in Sec. III B 2. To determine, for example, whether there was significant interaction between gender and ACT score, we implement the following model:

$$\begin{aligned}\text{Grade}_{ij} = {} & \gamma_{00} + u_{0j} + \gamma_{10}(\text{CWC ACT}_{ij}) \\ & + \gamma_{20}(\text{CWC Age}_{ij}) + \gamma_{30}(\text{Gender}_{ij}) \\ & + \gamma_{40}(\text{URM}_{ij}) + \gamma_{50}(\text{FG}_{ij}) + \gamma_{60}(\text{Citizen}_{ij}) \\ & + \gamma_{70}(\text{CWC ACT}_{ij})(\text{Gender}_{ij}) + r_{ij}.\end{aligned} \tag{4}$$

When interpreting the interaction term estimates, it is important to remember that the ACT score is mean centered, so the sign of the interaction effect changes from above to below the ACT mean. Only the demographic groups with statistically significant interaction terms are reported, and these models all had significantly better AIC than the same model without the interaction term.

Finally, when investigating demographic differences in performance in various components of the course grade in Sec. III C, multilevel modeling was used to account for student level factors and clustering within lecture sections. Using gender as the demographic factor, an example model equation using both gender and ACT to predict, say, homework grade was

$$\begin{aligned}\text{HW}_{ij} = {} & \gamma_{00} + u_{0j} + \gamma_{10}(\text{CWC ACT}_{ij}) \\ & + \gamma_{20}(\text{Gender}_{ij}) + r_{ij}\end{aligned} \tag{5}$$

Throughout analysis in each section, we use Nakagawa and Schielzeth's method for finding R^2 for multilevel models [29]. Since the goal is to quantify the total amount of variance explained by each model, we use the conditional R^2, denoted by $R_{2\text{GLMM}(c)}$, which provides the proportion of variance explained by both the random and fixed factors.

III. RESULTS

A. Mean demographic differences in performance

The descriptive table of mean ACT math scores and grade outcomes by demographics in Table I addresses our first research question and provides a general characterization of our local context. Overall, we find no significant difference in mean grade between male and female students in both algebra-based and calculus-based courses (however, we do find an important interaction with ACT score, as discussed below). This finding is consistent with the findings discussed in the introduction, namely that gender differences are varied but relatively small. Note also that no significant mean grade difference exists, even though there is a small but significant difference in ACT scores, with males scoring about 0.15 SD higher. In the algebra-based course, males had significantly higher DFW rates (18.7%) than females (14.6%), but males also received A's at a higher rate (36%) than female students (33.4%), indicating that males have a broader, flatter grade distribution. This difference in grade distribution was not observed in the calculus-based course.

Also consistent with prior literature, URM students and FG students receive significantly lower grades than their non-URM and non-FG counterparts, with mean differences of approximately 0.5 and 0.3 grade points, respectively. This translates to about 0.4 and 0.3 standard deviations. URM and FG students also have higher DFW rates, lower retention, and receive fewer A's than non-URM and non-FG students. However, it is also important to note differences in ACT score among these groups: URMs and first-generation students have scores about 0.7 standard deviations below their comparison groups. Therefore, it will be important to account for ACT score when considering the causes for demographic differences, as we will do in the next section.

There are some significant differences in retention rates among some of the demographic groups. URMs have about 10% lower retention in both courses, and FG students are about 8% lower only in the calculus-based course. Perhaps surprisingly, we found no significant difference in retention rates between men and women. The overall retention rates are of concern, especially for the calculus-based course, which is at about 70%. These rates seem to be low, though we do not have comparative data. What is an "acceptable" rate is not clear and should be subject to further inquiry, but beyond the scope of this paper. Of special concern is the very low rates of retention for students with no standardized test scores reported, which are typically students who have transferred from other institutions.

Statistically significant but relatively small differences in age exist between gender, URM status, first-generation status, and citizenship status groups. The only exception is the age difference between the students with and without an SAT or ACT score reported. This difference highlights the existence of a possibly underserved population of older students, who on average are scoring 0.3–0.4 grade points lower than the rest of the population. While this group is not part of the current focus of this paper, we speculate that age may act as another indicator for early (i.e., high school) preparation for the calculus-based introductory physics course. As most students take this course as a prerequisite for engineering requirements, students need to take this course early in their college careers. Thus, the younger the students are at the time of taking the course, the more likely they had the necessary preparation in high school. In the algebra-based course, the effect of age is less clear, but it may also be due to high school preparation. Many students enroll in this course later in their career since it is not a prerequisite for most of their major programs. Subsequently, students often put off taking the course until the end of students' college career, which is more temporally distant than their math and possibly physics preparation they completed in high school or their first few semesters in college.

Clear differences in performance between foreign and domestic students were found; as seen in Table I, foreign students earn a higher mean grade, lower DFW rates, and more A's than domestic students. In the algebra-based course, foreign students have a 15% lower retention rate, though it is not clear why.

Finally, about 10% of the population did not report ACT or SAT scores. This subpopulation is not proportionally representative of this whole population of students (6% and 7% more men, 13% and 12% fewer non-URMs, 6% and 3% fewer first generation students, and 20% and 19% more foreign students in the algebra- and calculus-based courses, respectively), and beyond the possibility that many are transfer students who tend to not report scores, it is not clear why they do not report standardized test scores. Certainly, this subpopulation warrants further investigation but that is beyond the scope of this paper.

B. Differences in grades controlling for ACT math, age, and lecture section

The summary table of descriptive statistics indicates some important differences in mean grades by demographic group, but there are also differences in mean ACT scores and ages. To what extent do differences in ACT scores and age "account" for differences in the grades? Essentially this is our second research question. To address this, we first look at the main effects of demographic group. Since we find, as others have found, that ACT is strongly associated with grade, we also explore the question as to whether there are any interaction effects between demographic group and ACT score.

1. Main effects of demographic group

Table I displays main-effect estimates of demographic differences in grade, DFW, A, and retention rates, using the models represented by Eqs. (1) and (2). These models

control for ACT score, age, and lecture section. The results indicate that ACT math, age, URM, first-generation status, and citizen status are all significant predictors of grade in the calculus-based course, and all but citizenship status are significant predictors of grade in the algebra-based course.

As expected, the model estimates that ACT score has a strong association with grade: increasing the ACT score by one point (about 0.25 SD) increases the grade by about 0.13 grade points, or about 0.1 SD. Age has a much weaker but negative effect on grades: increasing age by one year (0.5 SD) *decreases* the grade by 0.05 grade points. The effect of gender on grade is small or nonsignificant. If anything, controlling for ACT reveals that women have slightly higher grades (0.05) than men in the algebra-based course.

The estimated effect of URM and FG status on grades is also fairly small, of order 0.1 grade points. Notice that this magnitude of difference for URM and FG is smaller than the 0.3–0.5 grade point differences found in Table I. Thus, controlling for ACT scores substantially reduces these demographic differences, though they are still not zero: something more than just differences in ACT scores is causing differences in grades for URM and FG students. Finally, the model indicates that, after controlling for ACT, international students (who score almost 1 SD higher in ACT math) score 0.14 grade points higher than domestic students in the calculus-based course, but there is no difference in the algebra–based course.

One can determine the grade outcome probabilities by using Eqs. (2) and (3) and the logistic regression estimates in Table II. The logistic regression for the probability of receiving an A decreases significantly for URM and FG status but not for gender in both courses. For example, mean-aged and mean-ACT white male, non-first-generation students have a 0.31 and 0.17 probability of receiving an A in the algebra-based and calculus-based courses, respectively, while similar URM students have a 0.26 and 0.12 probability and similar FG students have a probability of 0.25 and 0.14 for receiving an A in those respective courses. In that same group of students, international students have a probability of 0.23 for receiving an A compared to domestic students in the calculus-based course.

Table II indicates a pattern for DFW probabilities that is consistent with the patterns for receiving an A. For mean-aged and mean-ACT white male, non-first-generation students, the probability of receiving a DFW is 0.10 and 0.12 for algebra-based and calculus-based students. For women it is 0.07 and 0.11, for URM it is 0.12 (not significant) and 0.17 and for FG it is 0.13 and 0.15, for algebra-based and calculus-based students, respectively.

The retention probabilities are also dependent on ACT scores, as shown in Table II and perhaps as expected. Even accounting for ACT score, FG students are still retained at a lower rate in the calculus-based course, and international (noncitizen) students are much less likely to be retained in the algebra-based course, though it is not clear why this would be the case.

In sum, we still see many demographic differences in grade outcomes, and when "controlling" for ACT scores, age, and lecture section, these differences are reduced for URM, FG and citizenship status, and are slightly increased in favor of women. Here, it is important to keep in mind that distributions of ACT scores differ by population, as seen for example in Table I.

TABLE II. Nonstandardized multilevel model fixed effects estimates with standard errors reported in parentheses. A significance of $p < 0.05$ is indicated by *. The random effect of lecture section and its standard deviation are reported in italics.

Predictor of grade	Algebra based	Calculus based
Intercept	2.81 (0.03)*	2.56 (0.02)*
CWC ACT math	0.131 (0.003)*	0.127 (0.003)*
CWC age	−0.05 (0.01)*	−0.03 (0.01)*
Gender	0.05 (0.02)*	0.04 (0.02)
URM	−0.11 (0.03)*	−0.19 (0.04)*
First generation status	−0.14 (0.03)*	−0.13 (0.03)*
Citizen status	−0.03 (0.04)	0.14 (0.03)*
Lecture section	$SD = 0.17$	$SD = 0.11$
Student-level residual	$SD = 0.91$	$SD = 0.93$
$R_{2\text{GLMM}(c)}$	0.28	0.21
Predictor of DFW		
Intercept	−2.20 (0.08)*	−1.95 (0.07)*
CWC ACT math	−0.26 (0.01)*	−0.23 (0.01)*
CWC Age	0.12 (0.03)*	0.11 (0.02)*
Gender	−0.37 (0.08)*	−0.19 (0.07)*
URM	0.18 (0.11)	0.39 (0.10)*
First generation status	0.33 (0.09)*	0.23 (0.08)*
Citizen status	−0.04 (0.17)	−0.10 (0.13)
Lecture section	$SD = 0.38$	$SD = 0.39$
$R_{2\text{GLMM}(c)}$	0.29	0.21
Predictor of A		
Intercept	−0.78 (0.07)*	−1.61 (0.06)*
CWC ACT math	0.28 (0.01)*	0.29 (0.01)*
CWC age	−0.04 (0.02)	0.02 (0.02)
Gender	−0.05 (0.06)	0.06 (0.07)
URM	−0.26 (0.10)*	−0.36 (0.12)*
First generation status	−0.33 (0.07)*	−0.24 (0.08)*
Citizen status	−0.01 (0.09)	0.42 (0.08)*
Lecture section	$SD = 0.35$	$SD = 0.32$
$R_{2\text{GLMM}(c)}$	0.31	0.26
Predictor of retention		
Intercept	1.51 (0.13)*	0.93 (0.07)*
CWC ACT math	0.11 (0.01)*	0.11 (0.01)*
CWC age	−0.18 (0.04)*	−0.03 (0.03)
Gender	−0.09 (0.10)	0.05 (0.08)
URM	−0.21 (0.14)	−0.12 (0.11)
First generation status	0.10 (0.12)	−0.24 (0.09)*
Citizen status	−0.86 (0.20)*	0.08 (0.13)
Lecture section	$SD = 0.68$	$SD = 0.42$
$R_{2\text{GLMM}(c)}$	0.19	0.09

TABLE III. Multilevel model results with significant interaction terms. Nonstandardized fixed effects estimates are reported with standard errors given in parentheses. A significance of $p < 0.05$ is indicated by *. The random effect of lecture section and its standard deviation are reported in italics.

	Algebra-based course		Calculus-based course	
Predictor of grade	Gender x ACT math	Citizen status x ACT math	Gender x ACT math	URM status x ACT math
Intercept		2.81 (0.03)*	2.56 (0.02)*	2.56 (0.02)*
CWC ACT math		0.134 (0.003)*	0.122 (0.004)*	0.124 (0.003)*
CWC age		−0.05 (0.01)*	−0.03 (0.01)*	−0.03 (0.01)*
Gender		0.05 (0.02)*	0.04 (0.02)	0.04 (0.02)
URM		−0.12 (0.04)*	−0.18 (0.04)*	−0.15 (0.04)*
First generation status		−0.14 (0.03)*	−0.13 (0.03)*	−0.13 (0.03)*
Citizen status		0.05 (0.05)	0.14 (0.03)*	0.15 (0.03)*
Interaction term		−0.03 (0.01)*	0.023 (0.007)*	0.03 (0.01)*
Lecture section		$SD = 0.17$	$SD = 0.11$	$SD = 0.11$
Student-level residual		$SD = 0.91$	$SD = 0.93$	$SD = 0.93$
$R_{2\text{GLMM}(c)}$		0.29	0.21	0.21
Predictor of DFW				
Intercept	−2.14 (0.09)*		−1.93 (0.07)*	
CWC ACT math	−0.23 (0.02)*		−0.22 (0.01)*	
CWC age	0.12 (0.03)*		0.11 (0.02)*	
Gender	−0.50 (0.09)*		−0.31 (0.09)*	
URM	0.17 (0.11)		0.38 (0.10)*	
First generation status	0.32 (0.09)*		0.23 (0.08)*	
Citizen status	−0.04 (0.17)		−0.10 (0.13)	
Interaction term	−0.06 (0.02)*		−0.07 (0.03)*	
Lecture section	$SD = 0.39$		$SD = 0.39$	
$R_{2\text{GLMM}(c)}$	0.30		0.22	
Predictor of A				
Intercept	−0.74 (0.07)*		−1.56 (0.06)*	−1.60 (0.06)*
CWC ACT math	0.25 (0.01)*		0.26 (0.01)*	0.28 (0.01)*
CWC age	−0.05 (0.02)*		0.04 (0.02)	0.02 (0.02)
Gender	−0.12 (0.06)*		−0.16 (0.08)	0.06 (0.07)
URM	−0.25 (0.1)*		−0.34 (0.12)*	−0.42 (0.13)*
First generation status	−0.33 (0.07)*		−0.24 (0.08)*	−0.24 (0.08)*
Citizen status	−0.02 (0.09)		0.42 (0.08)*	0.43 (0.08)*
Interaction term	0.06 (0.02)*		0.13 (0.02)*	0.10 (0.05)*
Lecture section	$SD = 0.36$		$SD = 0.32$	$SD = 0.32$
$R_{2\text{GLMM}(c)}$	0.31		0.27	0.27

2. Demographic effects conditional on ACT score

Because the grades depend so strongly on ACT score, it is natural to investigate whether the effects of demographic group are conditional on ACT score, or in other words, if there are interactions between demographic groups and ACT score when predicting grade outcomes. We modeled only one interaction at a time, using Eq. (3) for each demographic group.

Table III presents results for models with significant interaction terms. For the algebra-based course, there were no interactions with URM or FG status, but as seen in Fig. 1, there is a small crossover interaction of gender and ACT math score on the probability of receiving an A. For students with below average ACT math scores, males were more likely to receive an A. For students with above average ACT math scores, females were more likely to receive an A. Further, male students were more likely than female students to receive a DFW for all but the lowest ACT math scores. There was also a small negative interaction with international status and grade for the algebra-based course: international students with higher ACT scores tended to have slightly lower grades than domestic students and those with lower ACT scores tended to have slightly higher grades than domestic students.

For the calculus-based course, there were significant interactions for gender and URM status with ACT score. As seen in Fig. 2, there is a small crossover interaction effect of gender on the effect of ACT math on course grade. One of the most striking interaction results here is the interaction of URM status and ACT math on course grade. Only the highest ACT math score URM students receive the same grade as their non-URM counterparts; the

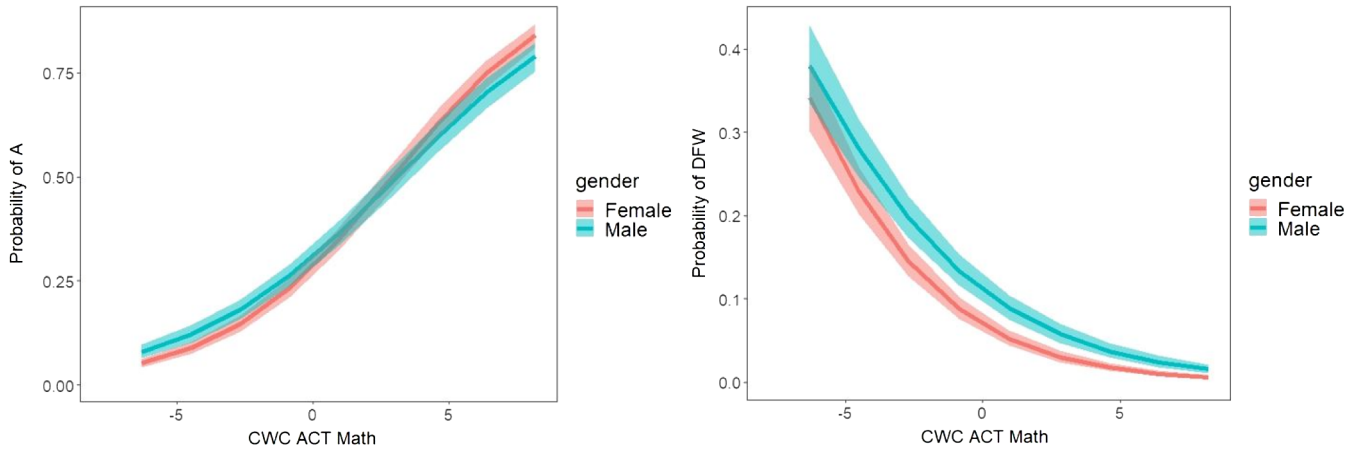

FIG. 1. Algebra-based course plots of significant interactions; probability of receiving an A vs group-centered ACT math score by gender and probability of receiving DFW vs group-centered ACT math score by gender.

majority of URM students earn lower course grades than non-URM students with the same ACT math scores. The lower the ACT math score, the more the URM grade gap widens.

C. Demographic differences in grade components

When examining grade outcomes in Table I, we observe gaps between demographic groups. However, as mentioned in the introduction, the gaps between demographic groups

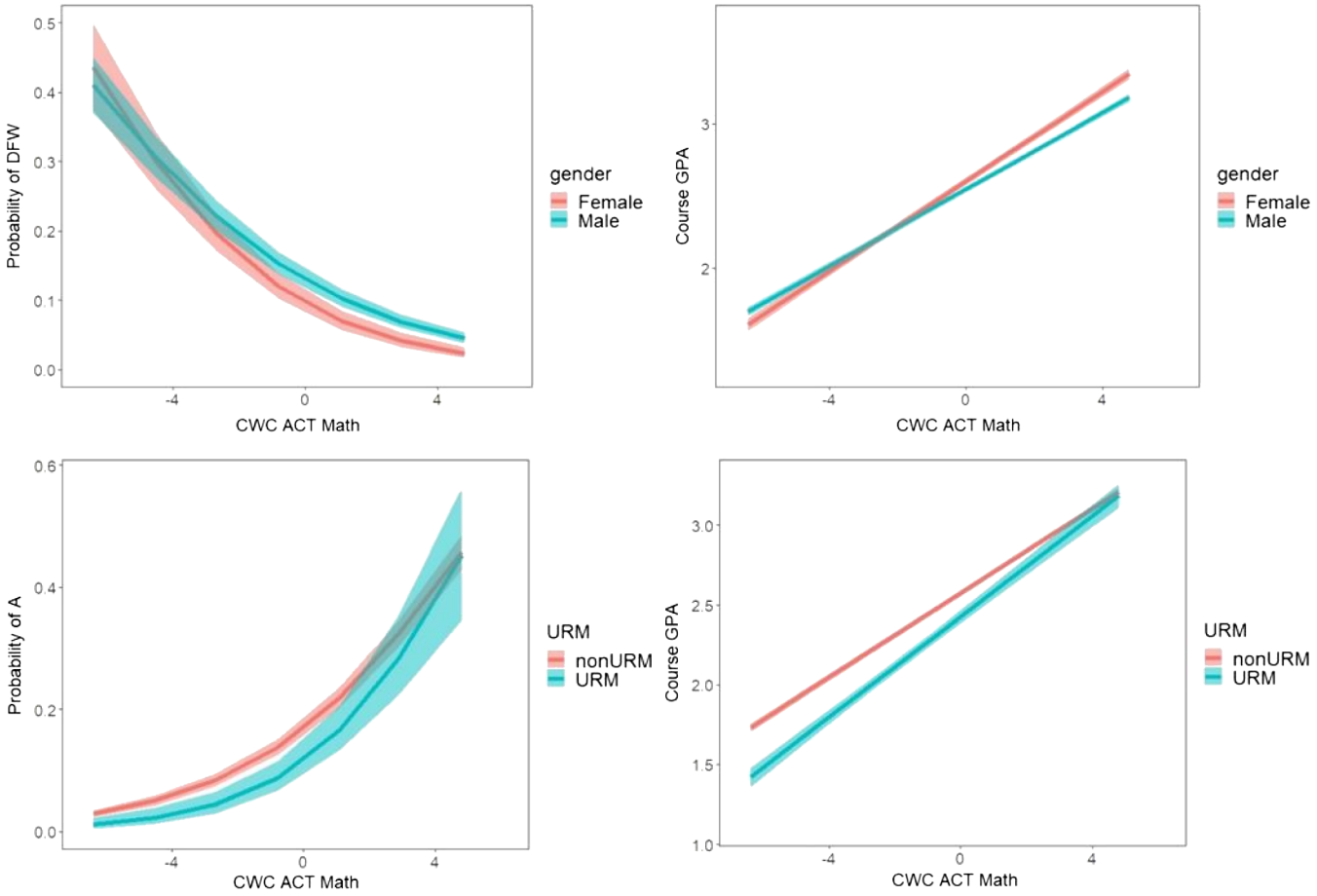

FIG. 2. Calculus-based course plots of significant interactions; probability of receiving a DFW vs group-centered ACT math score by gender, course GPA vs group-centered ACT math score by gender, probability of receiving an A vs group-centered ACT math score by URM status, and course GPA vs group-centered ACT Math score by URM status.

may depend on the components that comprise the grade [14]. To determine whether we could confirm these findings in our context and to further understand demographic differences in course performance, we investigated the extent to which all the components of grade: nonexam components (lab grade, homework scores) and test or exam components (in-class formal quiz, midterm exam and final exam scores) depended on demographic group and ACT score. We report the full results for all components in Tables X and XI in the Appendix, and a more summarized compact form for the total of the exam and nonexam components in Table IV.

Overall, for gender we found trends similar to previous findings: female students score higher on nonexam components than male students and about the same or slightly worse than male students on exam or test components. For example, in the calculus-based course there is a cross-over effect; women on average scored 0.24 standard deviations *higher* than male students on the nonexam components (labs, homework, and online essential skills practice

TABLE IV. Nonstandardized multilevel model fixed effects estimates with standard errors reported in parentheses. A significance of $p < 0.05$ is indicated by *. The random effect of lecture section and the student-level residual standard deviations are reported in italics. The maximum score on components is 100 (percent), and units of all numbers are in these units.

	Exam components		Nonexam components	
Predictor: Algebra-based course	$M = 76.0$	$SD = 17.5$	$M = 93.0$	$SD = 14.0$
Gender	$b_{\text{gender}} = -0.04\,(0.65)$ $SD_{\text{lec}} = 2.05$ $SD_{\text{lec}} = 17.37$ $R_{2\text{GLMM}(c)} = 0.014$		$b_{\text{gender}} = 4.26\,(0.51)^*$ $SD_{\text{lec}} = 1.42$ $SD_{\text{SLR}} = 13.76$ $R_{2\text{GLMM}(c)} = 0.033$	
Gender + Math ACT	$b_{\text{gender}} = 0.34\,(0.55)$ $b_{\text{Math}} = 2.23\,(0.07)^*$ $SD_{\text{lec}} = 2.00$ $SD_{\text{SLR}} = 14.04$ $R_{2\text{GLMM}(c)} = 0.291$		$b_{\text{gender}} = 4.08\,(0.51)^*$ $b_{\text{Math}} = 0.61\,(0.06)^*$ $SD_{\text{SLR}} = 1.23$ $SD_{\text{SLR}} = 12.90$ $R_{2\text{GLMM}(c)} = 0.060$	
URM	$b_{\text{URM}} = -8.88\,(0.98)^*$ $SD_{\text{lec}} = 2.16$ $SD_{\text{SLR}} = 17.16$ $R_{2\text{GLMM}(c)} = 0.042$		$b_{\text{URM}} = -2.95\,(0.80)^*$ $SD_{\text{lec}} = 1.73$ $SD_{\text{SLR}} = 13.95$ $R_{2\text{GLMM}(c)} = 0.020$	
URM + Math ACT	$b_{\text{URM}} = -2.63\,(0.86)^*$ $b_{\text{Math}} = 2.18\,(0.07)^*$ $SD_{\text{lec}} = 2.06$ $SD_{\text{SLR}} = 14.01$ $R_{2\text{GLMM}(c)} = 0.296$		$b_{\text{URM}} = -1.86\,(0.81)^*$ $b_{\text{Math}} = 0.54\,(0.07)^*$ $SD_{\text{lec}} = 1.47$ $SD_{\text{SLR}} = 13.08$ $R_{2\text{GLMM}(c)} = 0.043$	
FG	$b_{\text{FG}} = -5.87\,(0.75)^*$ $SD_{\text{lec}} = 2.05$ $SD_{\text{SLR}} = 17.20$ $R_{2\text{GLMM}(c)} = 0.034$		$b_{\text{FG}} = -1.23\,(0.60)^*$ $SD_{\text{lec}} = 1.66$ $SD_{\text{SLR}} = 13.91$ $R_{2\text{GLMM}(c)} = 0.015$	
FG + Math ACT	$b_{\text{FG}} = -2.40\,(0.65)$ $b_{\text{Math}} = 2.18\,(0.07)^*$ $SD_{\text{lec}} = 1.99$ $SD_{\text{SLR}} = 14.00$ $R_{2\text{GLMM}(c)} = 0.295$		$b_{\text{FG}} = -0.72\,(0.61)$ $b_{\text{Math}} = 0.56\,(0.06)^*$ $SD_{\text{lec}} = 1.43$ $SD_{\text{SLR}} = 13.03$ $R_{2\text{GLMM}(c)} = 0.041$	
	Exam components		Nonexam components	
Predictor: Calculus-based course	$M = 72.0$	$SD = 17.5$	$M = 91.4$	$SD = 13.6$
Gender	$b_{\text{gender}} = -2.14\,(0.66)^*$ $SD_{\text{lec}} = 3.73$ $SD_{\text{SLR}} = 17.16$ $R_{2\text{GLMM}(c)} = 0.048$		$b_{\text{gender}} = 3.28\,(0.51)^*$ $SD_{\text{lec}} = 1.77$ $SD_{\text{SLR}} = 13.41$ $R_{2\text{GLMM}(c)} = 0.028$	

(Table continued)

TABLE IV. *(Continued)*

	Exam components		Nonexam components	
Predictor: Calculus-based course	$M = 72.0$	$SD = 17.5$	$M = 91.4$	$SD = 13.6$
Gender+Math ACT	$b_{\text{gender}} = -0.96\,(0.48)$*		$b_{\text{gender}} = 3.36\,(0.50)$*	
	$b_{\text{Math}} = 2.31\,(0.08)$*		$b_{\text{Math}} = 0.58\,(0.07)$*	
	$SD_{\text{lec}} = 3.53$		$SD_{\text{lec}} = 1.57$	
	$SD_{\text{SLR}} = 14.54$		$SD_{\text{SLR}} = 12.64$	
	$R_{2\text{GLMM}(c)} = 0.252$		$R_{2\text{GLMM}(c)} = 0.046$	
URM	$b_{\text{URM}} = -9.32\,(0.97)$*		$b_{\text{URM}} = -2.37\,(0.77)$*	
	$SD_{\text{lec}} = 3.69$		$SD_{\text{lec}} = 1.62$	
	$SD_{\text{SLR}} = 16.91$		$SD_{\text{SLR}} = 13.47$	
	$R_{2\text{GLMM}(c)} = 0.068$		$R_{2\text{GLMM}(c)} = 0.017$	
URM + Math ACT	$b_{\text{URM}} = -3.75\,(0.91)$*		$b_{\text{URM}} = -1.28\,(0.80)$	
	$b_{\text{Math}} = 2.26\,(0.08)$*		$b_{\text{Math}} = 0.54\,(0.07)$*	
	$SD_{\text{lec}} = 3.52$		$SD_{\text{lec}} = 1.48$	
	$SD_{\text{SLR}} = 14.53$		$SD_{\text{SLR}} = 12.81$	
	$R_{2\text{GLMM}(c)} = 0.259$		$R_{2\text{GLMM}(c)} = 0.034$	
FG	$b_{\text{FG}} = -5.95\,(0.69)$*		$b_{\text{FG}} = -2.78\,(0.55)$*	
	$SD_{\text{lec}} = 3.73$		$SD_{\text{lec}} = 1.62$	
	$SD_{\text{SLR}} = 17.02$		$SD_{\text{SLR}} = 13.42$	
	$R_{2\text{GLMM}(c)} = 0.063$		$R_{2\text{GLMM}(c)} = 0.021$	
FG + Math ACT	$b_{\text{FG}} = -2.15\,(0.64)$		$b_{\text{FG}} = -1.79\,(0.56)$*	
	$b_{\text{Math}} = 2.27\,(0.08)$*		$b_{\text{Math}} = 0.51\,(0.07)$*	
	$SD_{\text{lec}} = 3.53$		$SD_{\text{lec}} = 1.40$	
	$SD_{\text{SLR}} = 14.51$		$SD_{\text{SLR}} = 12.69$	
	$R_{2\text{GLMM}(c)} = 0.256$		$R_{2\text{GLMM}(c)} = 0.034$	

assignments). However, for the exam components (quizzes, midterm 1, midterm 2, and final exams, respectively), women scored 0.12 SD *lower* than men. Note that when controlling for ACT score, these gaps remain.

Further, when considering differences between gender, it is important to note that ACT scores only accounted for about 2% of the variance in scores on nonexam components but a much larger portion—about 20%—of the variance for exam components. This change in variance explained can be determined by examining the $R_{2\text{GLMM}(c)}$ for each model.

Examination of Table IV (and Tables X and XI in the Appendix) reveals that both URM and FG status students tended to have lower scores on both exam and nonexam components, though the biggest differences were typically with the exam scores. Controlling for ACT score appreciably decreased the differences, especially for exam components. However, there was still a noticeable difference between nonexam and exam grade components: URM and FG students scored significantly worse on the exam components. Further, similar to the case for gender, for URM and FG students ACT scores only accounted for 1%–2% of the variance for nonexam components, but typically accounted for about 20% for exam components.

Note also that, comparing our results to the results from Salehi *et al.* [30] for final exam scores in physics courses at a large midwestern public research university (similar to OSU), we found most of our results to be similar to theirs, and a couple that appear to be different. The differences, though statistically significant, are not large (in effect size) and may simply reflect relatively minor variations in local contexts. It is worth noting that one clear agreement between our study and Salehi *et al.* was that the ACT scores were moderately well associated with final exams scores: ACT scores explained about 13%–15% of the variance in final exam scores in their dataset.

In sum, different demographic groups perform differently on exam vs nonexam grade components, with minority students typically performing worse on exam components. Further, the exam components are moderately associated with ACT scores, but the nonexam components are not.

D. Grade outcomes using different grading models

Since demographic performance gaps depend on the grade component, it is reasonable to consider that changing the grade component weights could affect different demographic groups in different ways. Specifically, reducing the weight of exam components could differentially benefit women, URM, and FG students. Put another way, the current weighting may be disadvantaging these groups. Further, the common weighting of components strongly

favors exam components, and these components are more strongly correlated with ACT score than nonexam components. Thus, it follows that reducing the weighting of exam components could also reduce the correlation of final grade with ACT scores. One caveat to this "thought experiment" is that shifting the grade weights could significantly influence student behavior and performance as well as instructor behavior and construction of the course components themselves. We will discuss this issue more in Secs. IV. C and IV. D, but for now we will assume that, with relatively small shifts in the weights, there will be relatively small shifts in behavior and even smaller relative differences in behavior among demographic groups.

In order to determine what could have been the effect of a modification of grade weights on the grade outcomes on women, URM, and FG students, we use the original student grade component data to counterfactually model a decrease in the exam component weight. Specifically, we compare four different grading models: two grade components weighting (original weight vs lower exam score weight) crossed with two different mean course grades (original mean grade vs higher mean grade). The four models are outlined in Table V. Model 0 is the actual original, unmodified grade component data. Model III changes the grade component weighting according to Table VI: the weight of exam components is reduced from 70% to 50% with a corresponding increase in the non-exam components. We chose these weightings as a relatively small and potentially feasible (and departmentally acceptable) change from the original weighting. Since students typically have higher scores in non-exam components (see Table IV and Tables X and XI in the Appendix) changing the weights in model III also inadvertently increases the mean grade for the course. In order to compare different weightings but with identical mean grades, we introduced model I which linearly adjusts all of the grade cutoffs [translating total points (such as 85%) to grade points (such as 3.3)] such that the mean grade is the same as the original mean. Finally, to compare to a simple increase in mean grades, we introduced model II which maintains the original grade component weighting, but linearly adjusts all of the grade cutoffs such that the mean grade is the same as model III.

1. Comparison of mean grade outcomes

The results are presented in Table VII. Note that the model 0 outcomes are slightly different than the outcomes in Table I. This is because Table VII represents only a subset (2 years) of the data in Table I (6 years), since we only were able to obtain grade component data for this subset. Also note that for whole-course mean grades, the difference between the original mean and "higher mean grade" was +0.23 grade points and the median changed from a B to a $B+$ for the algebra-based course, and for the calculus-based course, the overall course mean shifted by +0.28 grade points, and the median remained constant at a B.

There are several noteworthy outcomes from the grade weight models. Most notably, as expected from the results of Table IV, the four grading models have different grade outcomes for different demographic groups. Let us consider the mean GPA outcomes first. Figure 3 graphically presents differences in mean grades between a given demographic group and its comparison group. The error bars were calculated using nonparametric bootstrapping.

Overall, in most cases women, URMs, FGs, and foreign students differentially benefit from the modified grade weighting, namely, models I and III. For example, for the calculus-based course, In model 0 women have a 0.09 lower mean grade than men while in models I and III they have the same mean as men. Further, in model 0 URMs have a 0.60 lower mean grade than non-URMs, but in model III this difference is decreased to 0.49. These differences between models are somewhat small in size but statistically significant ($p < 0.05$).

Table VII also shows that the DF and A rates also vary significantly among models. This is to be expected since the weighting to nonexam components or higher mean-grade shift both increased grades. What is more interesting is that the *relative* differences in the DF and A rates among the models and demographic groups are dramatic in some cases. The relative percent changes compared to model 0 can be difficult to discern in Table VII, so we produced Figs. 4 and 5 to make these changes more visible. Figure 4 presents the decrease in percentage of D's or F's given to students compared to model 0. For example, in the calculus-based course, about 22% fewer females, compared

TABLE V. The four grading models.

	Original weighting	*Modified weighting*
Original mean grade	0	I
Higher mean grade	II	III

TABLE VI. Original percentage of grade components used to calculate the grades students received vs modified percentage of grade components used as a comparison model.

Grade component	*Original percentage of grade*	*Modified percentage of grade*
Online homework	15	25
Essential skills	1	5
Lab	14	20
Quizzes	15	7.5
Midterm 1	15	7.5
Midterm 2	15	10
Final Exam	25	25
Total exam component	*70*	*50*

TABLE VII. Grade, % of students receiving D or F, and % of students receiving an A by course, demographic, and grade model. The subscripts 0, I, II, and III indicate the grade model used. Actual grade outcomes were compared to their modified versions and are listed in bold if a t test showed significant difference from model 0.

				$Grade_0$		$Grade_I$		$Grade_{II}$		$Grade_{III}$									
		Population	N	M	SD	M	SD	M	SD	M	SD	% DF_0	% DF_I	% DF_{II}	% DF_{III}	% A_0	% A_I	% A_{II}	% A_{III}
First semester algebra-based physics		Full population	2924	2.80	1.1	2.80	1.1	**3.03**	1.0	**3.03**	1.0	13.7	**11.7**	**10.9**	**9.4**	33.9	**30.2**	**44.7**	**42.5**
	Gender	Male	1183	2.79	1.2	2.72	1.1	**3.02**	1.1	**2.94**	1.1	15.5	15.0	**12.5**	**12.1**	35.4	**29.8**	**45.7**	**41.3**
		Female	1738	2.81	1.1	2.86	1.0	**3.05**	1.0	**3.10**	1.0	12.5	**9.3**	**9.7**	**7.6**	33.0	30.5	**44.1**	**43.4**
	URM status	URM	331	2.22	1.2	2.29	1.1	**2.48**	1.2	**2.52**	1.1	28.7	26.1	23.9	**20.3**	17.4	15.2	**26.8**	**23.9**
		nonURM	2492	2.80	1.1	2.81	1.0	**3.03**	1.1	**3.03**	1.0	13.5	**11.7**	**10.5**	**9.4**	34.1	**30.1**	**44.8**	**42.5**
	First generation status	FirstGen	657	2.48	1.2	2.54	1.1	**2.73**	1.2	**2.77**	1.1	21.2	18.0	**16.6**	**14.1**	24.9	22.4	**33.1**	**32.0**
		NonFirstGen	2267	2.89	1.1	2.88	1.0	**3.12**	1.0	**3.11**	1.0	11.6	9.8	**9.2**	**8.1**	36.5	**32.4**	**48.1**	**45.6**
	Citizen status	Foreign	393	3.16	1.0	3.08	0.9	**3.36**	0.9	3.29	0.9	6.6	6.6	5.3	5.8	49.5	43.4	**57.9**	54.8
		Domestic	2531	2.75	1.1	2.76	1.1	**2.98**	1.1	**2.99**	1.0	14.8	**12.4**	**11.7**	**10.0**	31.5	**28.2**	**42.7**	**40.6**
	No standardized test score reported		258	2.28	1.3	2.32	1.2	**2.53**	1.3	**2.54**	1.2	29.0	25.4	23.8	**20.6**	25.0	21.8	29.8	28.2
First semester calculus-based physics		Full population	3597	2.64	1.1	2.64	1.0	**2.92**	1.1	**2.92**	1.0	14.3	12.7	**10.6**	**9.7**	24.5	**19.5**	**36.9**	**33.9**
	Gender	Male	2738	2.67	1.1	2.64	1.0	**2.94**	1.1	**2.92**	1.0	13.9	12.8	**10.3**	**10.0**	25.6	**20.1**	**37.4**	**34.6**
		Female	845	2.56	1.1	2.64	1.0	**2.87**	1.0	**2.92**	1.0	15.3	12.2	**11.3**	**8.7**	21.4	17.9	**35.5**	**32.2**
	URM status	URM	304	2.09	1.2	2.17	1.1	**2.37**	1.2	**2.47**	1.1	29.2	24.2	23.1	**18.8**	10.1	6.5	**19.1**	**18.4**
		nonURM	3108	2.69	1.1	2.68	1.0	**2.97**	1.0	**2.96**	1.0	12.8	11.6	**9.5**	**8.8**	25.5	**20.4**	**38.4**	**35.1**
	First generation status	FirstGen	388	2.30	1.2	2.35	1.1	**2.59**	1.2	**2.60**	1.1	23.9	20.7	**17.6**	**16.2**	18.1	**14.0**	**27.1**	**24.1**
		NonFirstGen	2909	2.72	1.0	2.71	1.0	**3.00**	1.0	**2.99**	1.0	12.0	10.8	**8.9**	**8.2**	26.0	**20.7**	**39.2**	**36.3**
	Citizen status	Foreign	539	3.10	1.0	3.02	1.0	**3.33**	0.9	**3.26**	0.9	6.9	7.5	5.4	5.8	43.9	37.9	**56.5**	**51.9**
		Domestic	3058	2.56	1.1	2.57	1.0	**2.85**	1.1	**2.86**	1.0	15.6	**13.6**	**11.5**	**10.4**	21.1	**16.2**	**33.4**	**30.8**
	No standardized test score reported		357	2.45	1.4	2.44	1.3	**2.70**	1.3	**2.67**	1.3	25.5	21.8	19.9	**18.8**	29.9	24.3	**38.3**	35.7

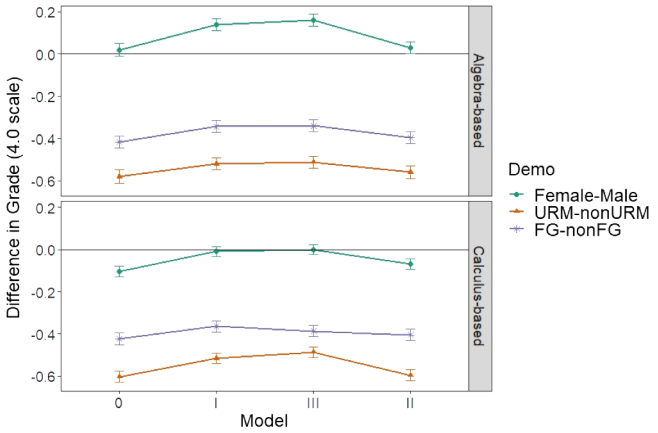

FIG. 3. Demographic differences in grade on a 4.0 scale for each model presented in Table V. The four grading models.

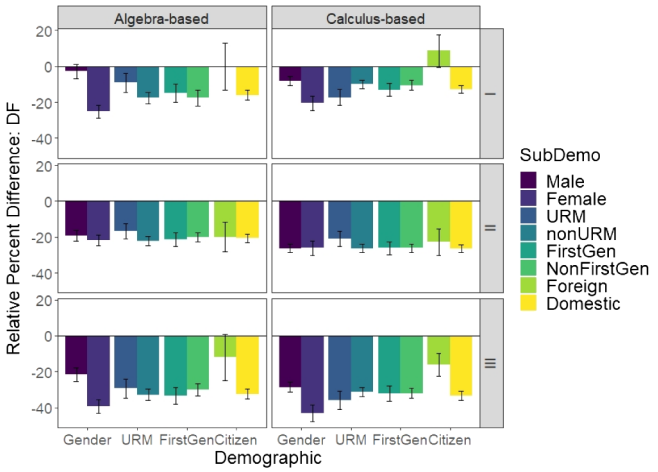

FIG. 4. Relative percent differences in D's and F's between model 0 and models I, II, or III, by demographic group and course. For example, if model I were used instead of model 0 in the algebra-based course, about 2% fewer males and about 25% fewer females would have received a D or F.

to 6% fewer males, would have received a D or F grade if model I were used, compared to what they actually received in model 0. For model I there is also an 18% drop in DF's for URMs compared to an 8% drop for non-URMs compared to model 0. For model III, there is 42% drop in DFs for women and 26% drop for men compared to model 0. Note that there are no differential DF rate effects for FG versus nonFG students.

There are corresponding differential benefits from the grade reweighting for women and URMs in receiving an A as well. For example, in model III, 50% more women and 35% more men would receive an A compared to model 0, and a dramatic 75% more URMs compared to 40% more non-URMs would receive an A. The increase in A's in models II and III is understandable given the mean grade is higher for these two models. Note that Model I has

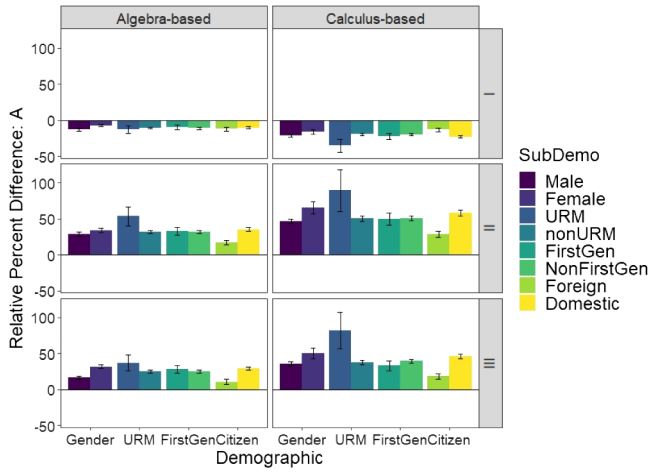

FIG. 5. Relative percent differences in A's between model 0 and models I, II, or III, by demographic group and course. For example, if model II were used instead of model 0 in the algebra-based course, about 50% more URMs and about 30% more non-URMS would have received an A.

fewer A's and fewer DFs, indicating a more narrow grade distribution for this model.

Finally, the grade-weighting models also show that students who are noncitizens tend to fare differentially worse in the modified grade weight. This is largely due to the fact that noncitizen students tend to score higher on exam grade components.

2. Grade weight models: demographic differences and ACT scores

Similar to Sec. III B 1 we can also model demographic differences for the four grade models including effects of ACT scores and random effects of lecture sections. We have computed such models and found two main results. The first is that, except for the estimates of the coefficient for the ACT scores, there were no statistically significant differences in the estimates of the regression coefficients compared to those found in Table II.

The second result, as summarized in Table VIII, is that the modified grade weight models reduced the dependency of grade on ACT score. Specifically, the ACT coefficient for the standard grade weight (model 0 and model II) was about 0.13 grade points per ACT point, while for the modified weighting (Models I and III) the ACT coefficient was less than 0.11 grade points for the ACT point. Mathematically speaking, this difference in slope translates to a difference in grade performance among demographic groups. For example, for the calculus-based course URMs tend to score 2.5 ACT points lower than non-URMs (see Table I). The change in ACT coefficient for model 0 ($b_{\mathrm{ACT}} \sim 0.13$) to model I or III ($b_{\mathrm{ACT}} \sim 0.10$) results in an average reduction of the gap between URMs and non-URMs by about 0.07 grade points.

TABLE VIII. Conditional R^2 and regression coefficients of ACT Math using models outlined in Eqs. (1) and (2) with ACT score and without ACT score as a predictor, both times including demographic factors (DEMO = gender, URM, FG, Citizen status, and age) as predictors and lecture section as random effect). The regression coefficients are not normalized, and the R^2 values reported are conditional R^2, which describes the proportion of variance explained by both the fixed and random factors.

		Algebra based			
Predictor of grade		Model 0	Model I	Model II	Model III
DEMO	$R_{2\mathrm{GLMM}(c)}$	0.14	0.13	0.13	0.12
DEMO + Math ACT	b_{Math}	0.135 (0.005)*	0.111 (0.005)*	0.127 (0.005)*	0.107 (0.005)*
	$R_{2\mathrm{GLMM}(c)}$	0.31	0.25	0.20	0.17
Predictor of DF					
DEMO	$R_{2\mathrm{GLMM}(c)}$	0.14	0.14	0.12	0.12
DEMO + Math ACT	b_{Math}	−0.30 (0.02)*	−0.23 (0.02)*	−0.29 (0.02)*	−0.20 (0.02)*
	$R_{2\mathrm{GLMM}(c)}$	0.36	0.28	0.33	0.23
Predictor of A					
DEMO	$R_{2\mathrm{GLMM}(c)}$	0.11	0.10	0.12	0.11
DEMO + Math ACT	b_{Math}	0.27 (0.02)*	0.25 (0.02)*	0.27 (0.02)*	0.24 (0.02)*
	$R_{2\mathrm{GLMM}(c)}$	0.29	0.25	0.28	0.25
		Calculus based			
Predictor of grade		Model 0	Model I	Model II	Model III
DEMO	$R_{2\mathrm{GLMM}(c)}$	0.10	0.09	0.09	0.09
DEMO + Math ACT	b_{Math}	0.129 (0.006)*	0.106 (0.005)*	0.122 (0.006)*	0.103 (0.005)*
	$R_{2\mathrm{GLMM}(c)}$	0.22	0.18	0.20	0.17
Predictor of DF					
DEMO	$R_{2\mathrm{GLMM}(c)}$	0.19	0.17	0.17	0.17
DEMO + Math ACT	b_{Math}	−0.25 (0.02)*	−0.21 (0.02)*	−0.26 (0.02)*	−0.21 (0.02)*
	$R_{2\mathrm{GLMM}(c)}$	0.30	0.26	0.30	0.26
Predictor of A					
DEMO	$R_{2\mathrm{GLMM}(c)}$	0.10	0.12	0.11	0.08
DEMO + Math ACT	b_{Math}	0.29 (0.02)*	0.29 (0.02)*	0.27 (0.02)*	0.24 (0.02)*
	$R_{2\mathrm{GLMM}(c)}$	0.26	0.26	0.24	0.20

Another way to consider the diminished influence of ACT score on the modified grade weight is to compare (subtract) the R-squared values for the models including versus not including the ACT score as a predictor. For the calculus-based course, the R-squared value attributable to the ACT score for models I and III (∼0.08) is lower than that for model 0 (∼0.12). This diminished influence of ACT score is expected, since the modified grade weight more strongly weights grade components that are not strongly correlated with ACT score.

IV. CONCLUSIONS AND DISCUSSION

Let us conclude by first summarizing our findings in terms of grades, grade components, and grade models, then we will discuss broader implications about ACT scores, grades, and grading policies.

A. Grades

We have investigated the grades for over 20 000 students in algebra-based and calculus-based introductory physics courses at our large, public research university and found that the mean grade differences between gender, URM, and FG status are consistent with findings in previous studies: there are small and nonsignificant differences between women and men (<0.03 SD) and significantly lower mean grades for URM students (∼0.4 SD) and FG students (∼0.3 SD). It is important to note, however, that there was a gender by ACT score interaction when predicting grade such that even though the mean grades did not vary by gender, women with high ACT scores had significantly higher grades than men, and women with low ACT had significantly lower grades than men. There was URM status by ACT score interaction for grades as well. We also found that significantly higher grades for foreign vs domestic students (∼0.4 SD). There were no significant differences in retention between genders, but retention of URM students was about 10 percentage points lower than non-URM students in both courses, and FG student retention was also about 10 percentage points lower in the calculus-based course.

We found, as expected, that ACT scores are moderately strongly correlated with grades and that there are differences in mean scores, with women scoring about 0.1 SD lower, URM students scoring about 0.7 SD lower,

and FG students scoring about 0.6 SD lower. Further, we found that age is negatively correlated with grade and that men, URM students, and non-FG students tend to be 0.2–0.3 years older.

When ACT score, age, and the random effects of lecture section were included in multilevel model predicting grade, we found that URM, FG, and citizenship status remain significant factors, though when controlling for ACT score and age, the grade gap for URM and FG students is cut roughly in half. It should also be noted that, even accounting for ACT score, URM, and FG students had significantly higher DFW rates. Further, there were some notable interactions; for example, there was a small interaction between gender and ACT score in predicting course grade with women with high ACT scores achieving higher grades than men with high ACT scores, and there was in interaction between URM status and ACT scores, with the grade gap at zero for the highest ACT scores, but widening as ACT score decreased.

It is important to keep in mind that this model is additive in intersectionality: for example, if a student is both a URM and FG, then each effect is added separately, compounding the gap.

B. Grade components

There were two important findings for grade components. First, we replicated and confirmed previous findings that women tend to do better than men on nonexam components (+0.3 SD) and the same or worse on exam components (−0.1 SD). In contrast to previous results, we also find a similar differential bias between grade components for URM and FG students: The gap tend to be smaller for nonexam components (−0.1 to −0.2 SD) than for exam components (−0.3 to −0.5 SD).

The second important result, consistent with previous findings as well, is that exam components tend to be more highly correlated with ACT scores than nonexam components. In other words, ACT scores typically account for about 20% of the variance of exam components but only about 2% of the variance of nonexam components.

C. Grade models

Given that performance on exam vs nonexam grade components varies by demographic group, it is possible that changing the weighting of the exam components could have differentially changed the mean grades for each demographic group, and, with the simplifying assumption that students and instructors would not have significantly changed their performance with the different weighting, this is indeed what we found. The reason for explicit modeling was to provide insight into how large the differences might be. The reduction in the exam component weight did tend to differentially benefit the minority demographic groups. For URM and FG students the mean grade gap was closed by a fairly small amount (at most 0.1 grade points). For women, the gap was eliminated in the calculus course, and for the algebra-based course, women performed 0.1–0.2 grade points better than men. It is interesting to note that from Table VIII and Fig. 3 that "simply raising the mean grade" with no changes to the grade weights (i.e., model II), does *not* help to close the gap between demographic groups, though it does differentially affect grade A rates.

In contrast to mean grade shifts, the relative changes in the percent of demographic groups receiving a DF or A was dramatic in some cases. For example, comparing model III with model 0 (the actual model), 42% fewer women vs 26% fewer men would have received a DF, and 75% more URM students vs 40% more non-URM students would have received an A.

An additional important outcome of the grade models was that the dependence of final grade on ACT score decreased when the weight of the exam components was decreased. This follows naturally, since the exam components were moderately corrected with ACT score while the nonexam components were not.

As mentioned earlier, an important caveat to these results is that a shift in the grade weights may also change student behavior and performance, and this could modify the results of the simple models used in this paper. But this caveat does not invalidate the general point of the findings in the paper, namely, that shifting grade weights may have a significant impact on achieving grade equity, and thus must be more explicitly considered. Further, whether or exactly how student behavior and performance would change if the weighting were changed is an empirical question that clearly warrants further investigation. In fact, there are credible reasons to believe that decreasing the exam component grade weight could also facilitate more equitable student performance in addition to the purely mathematical shifts shown in this paper. For example, test anxiety has been documented to influence exam performance for women more than men in several STEM areas [24], and decreasing the weight of exams has been shown to reduce the gap between men and women in biology classes [31–32].

Shifting grade weights may also prompt changes in instructional materials and methods, such as the content and nature of the newly weighted course components themselves. This will be discussed more at the end of the next section, which begins with a general discussion of issues regarding ACT scores.

D. Commentary and implications for grades and grading

ACT scores are often used by researchers as proxies for preparation or ability. As such, they are used to "control" for these student level characteristics. We use *control* in quotation marks because it is important to keep in mind that while it is an empirical fact that grades are correlated with

ACT scores and distributions of ACT scores differ by population, the extent to which this may, at least in part, "explain" differences in grade depends on what is assumed about what ACT scores and grades are measuring. We found that "even after controlling for ACT scores" there were differences for URM and FG students. One could ask why there is a remaining difference, and that is an important questions to ask, but the fact that "controlling" for ACT scores reduces demographic differences—even if it reduced the differences to zero—should only be considered as, at best, a partial mitigation or explanation of the differences in grades. Instead, the dependence itself of grades on ACT scores could also be a symptom of more systemic problems with our assessments and grading practices.

Yet ACT scores currently do empirically have predictive power for physics grades. We emphasize here that grades are correlated with ACT because the *exam* components—and only exam components—are appreciably correlated with ACT in our study. Consider Table IX, determined from our dataset, which shows the correlations among exam components, nonexam components, and ACT score. As stated earlier, it is clear that ACT scores are correlated with exam components and at best only weakly with non-exam components. This is perhaps not surprising: ACT tests, perhaps by design, are administered in a very similar format and context to traditional physics tests or exams. They are in sequestered, time limited venues, typically multiple choice, and high stakes. This context is somewhat different than the context for nonexam components, namely, homework and labs.

Table IX has one more important result to consider: nonexam components are moderately to strongly correlated with exam components. This implies that there is some set of skills and knowledge measured by the nonexam components that is independent of the ACT yet is directly related to performance on the exams. Put another way, if high performance on exams is something that is valued by the instructor, then this valued attribute is at least partially measured by the nonexam components of the course. In fact, it measures this attribute better than ACT scores. We performed a multiple regression on the data in Table IX and found that nonexam scores accounted for about 30% of the variance in exam scores, and ACT scores separately account for only about 20% of the variance.

The moderately strong dependence of exam components on ACT scores could be viewed as problematic: given that there are concerns that using ACT scores for admission into universities is contributing to the relatively low diversity of students accepted to the extent that many institutions are moving away from using ACT scores [33], are grades in college courses simply doing the same thing with current grading practices? This question is especially salient given that students use grades to make decisions whether to stay in STEM and oversubscribed STEM major programs use minimum grades as filters into their programs (cf. Ref. [20]). Is this at least partially equivalent to simply using ACT scores as cutoffs? Put another way, it appears as though at least to some extent traditional exams are filtering out diversity like ACT scores are filtering out diversity. From this perspective, exams and ACT score are two sides of the same problematic issue: using traditional tests to measure and award achievement.

TABLE IX. Correlations for the calculus-based course. The results are very similar for the non-calculus-based course.

	ACT score	Nonexam Comp
Exam components	0.46	0.61
Nonexam components	0.15	⋯

The nonexam components, however, are worthy of more consideration for at least three reasons. First, they have the virtues of not depending on ACT score and having smaller demographic biases, while still being strongly correlated with exam scores. This suggests that the nonexam components are measuring physics knowledge and skills as typically assessed by exams, yet do not suffer from the same demographic biases as the exam components. It is not clear why different demographic differences occur in exams vs nonexam components, though known effects of stereotype threat and stress may be strong candidates. It is certainly an interesting topic for further investigation. Second, it is clear that successful completion of nonexam components themselves, such as labs, homework, and in-class group work, are highly valued by instructors. These components are better suited for developing and assessing different and perhaps more important attributes of students valued by the physics community [34] that extend beyond physics knowledge to scientific, communication, and professional skills: working in groups, managing time, consistent weekly work, handling more difficult problems (that cannot be done in exam setting), using a variety of resources, etc. Third, shifting grade weights to such nonexam components may also send a positive message to students about the importance of these attributes.

These are important arguments for reconsidering grade weights and, by extension, grading practices in general. More specifically, we see two interrelated issues here. The first is that the instructor (or the entity responsible for grading policy) must determine what is an acceptable mean grade and grade distribution. A more detailed discussion of this issue, such as why we use grade curving rather than fixed standards, goes beyond the scope of this paper, but, as mentioned in the introduction, students (and programs) use grades to make choices about majors. There appears to be differences among demographic groups in how grades are interpreted, thus, the decision for an instructor to choose a given grade distribution is inevitably linked to the issue of demographic representation. Naturally, this issue could be mitigated by reducing the amount that grades factor into

admissions into programs. We bring up the issue of grade distributions because since grade components typically have different distributions themselves, changing the weights changes the grade distributions.

That leads to the second issue about grading policy. The increased attention, and possibly weight, to nonexam components leads us to reconsider how nonexam component grades are awarded and what they should award. For example, for the lab component, our department currently awards full points for reasonable group-work participation and a reasonable, though somewhat minimal, write-up. As a consequence, most students receive full credit for good faith efforts and performance. However, if more weight is given to the lab component, then should we reexamine how points are awarded, for example, requiring a higher level of performance? Consider also the homework component, which was online. If more weight is given to this, then this will likely provide incentive for more students to misuse available resources to complete the homework assignment (e.g., using online answering sites). Given this, how do we change the homework task to prevent students from completing it in unintended ways? We have recently made a change to have some component of the homework be hand-in, show-your-work format using specific rubrics for problem solving with the idea that this will increase student engagement—and allow for richer assessment of their work—in ways that are more aligned with our instructional goals.

All told, we will not know exactly how shifting weights will affect grade equity until it is tried, and we hope that this study can help to motivate instructors and departments to consider it. This brings us to our final point. A question that may help to guide an instructor or program to construct beneficial grading policies is as follows: are we awarding what we value? Both sides of this "equation" warrant deep consideration. On one side is what we as instructors or as a program must decide what we value and prioritize among goals such as physics knowledge, physics ways of thinking, problem solving and reasoning skills, professional skills, people skills, ethics, equity and inclusion. On the other side is a careful look at what we are awarding. Besides considering the knowledge, skills, and ways of thinking we value for our students, we must acknowledge that what we are awarding is inevitably tied to which students we are rewarding. In support of the views and empirical documentation in the literature discussed earlier, we find that current grades and grading may not be awarding what we value, and we should continue to investigate this issue and adopt policies and practices that bring us closer to balance.

ACKNOWLEDGMENTS

Funding for this research was provided by the Center for Emergent Materials: an NSF Materials Research Science and Engineering Centers (MRSEC) under Grant No. DMR-1420451. We also gratefully acknowledge the cooperation of Tom Gramila and the rest of the physics department at Ohio State University, and especially Tom Barrett for his valuable assistance in retrieving the registrar data used in this study.

APPENDIX

Disparities in grade between demographic groups may depend on individual grade components, as discussed in Sec. III.C. To investigate the extent to which specific grade components depend on ACT score and various demographic variables, multilevel modeling was performed according to the example equation given in Eq. (5). Table IV presents a compact summary of the results; the full results for each individual grade component for both the algebra- and calculus-based courses are presented in this appendix in Tables X and XI.

TABLE X. Means, standard deviations, regression coefficients, and conditional R_2 values for multiple regression models using demographics and math ACT scores to predict course grade components in the algebra-based course.

Predictor: Algebra-based course	Lab	HW	ES	Quizzes	Midterm 1	Midterm 2	Final exam	Total Pts
	$M = 97.1$ $SD = 11.3$	$M = 89.9$ $SD = 19.1$	$M = 81.1$ $SD = 29.8$	$M = 80.6$ $SD = 16.5$	$M = 74.8$ $SD = 19.4$	$M = 74.5$ $SD = 21.5$	$M = 74.8$ $SD = 20.6$	$M = 81.3$ $SD = 15.3$
Gender	$b_{\text{gender}} = 2.20\ (0.42)*$ $SD_{\text{lec}} = 0.84$ $SD_{\text{SLR}} = 11.19$ $R_{2\text{GLMM}(c)} = 0.015$	$b_{\text{gender}} = 5.64\ (0.69)*$ $SD_{\text{lec}} = 2.19$ $SD_{\text{SLR}} = 18.67$ $R_{2\text{GLMM}(c)} = 0.035$	$b_{\text{gender}} = 12.17\ (1.06)*$ $SD_{\text{lec}} = 6.26$ $SD_{\text{SLR}} = 28.51$ $R_{2\text{GLMM}(c)} = 0.084$	$b_{\text{gender}} = 0.25\ (0.60)$ $SD_{\text{lec}} = 3.90$ $SD_{\text{SLR}} = 16.06$ $R_{2\text{GLMM}(c)} = 0.056$	$b_{\text{gender}} = -1.74\ (0.72)*$ $SD_{\text{lec}} = 1.48$ $SD_{\text{SLR}} = 19.32$ $R_{2\text{GLMM}(c)} = 0.008$	$b_{\text{gender}} = 1.29\ (0.80)$ $SD_{\text{lec}} = 1.40$ $SD_{\text{SLR}} = 21.48$ $R_{2\text{GLMM}(c)} = 0.005$	$b_{\text{gender}} = 0.01\ (0.76)$ $SD_{\text{lec}} = 3.05$ $SD_{\text{SLR}} = 20.34$ $R_{2\text{GLMM}(c)} = 0.022$	$b_{\text{gender}} = 1.07\ (0.56)*$ $SD_{\text{lec}} = 1.83$ $SD_{\text{SLR}} = 15.16$ $R_{2\text{GLMM}(c)} = 0.015$
Math ACT + Gender	$b_{\text{Math}} = 0.26\ (0.05)*$ $b_{\text{gender}} = 1.90\ (0.42)*$ $SD_{\text{lec}} = 0.56$ $SD_{\text{SLR}} = 10.64$ $R_{2\text{GLMM}(c)} = 0.018$	$b_{\text{Math}} = 0.90\ (0.09)*$ $b_{\text{gender}} = 5.60\ (0.69)*$ $SD_{\text{lec}} = 1.99$ $SD_{\text{SLR}} = 17.54$ $R_{2\text{GLMM}(c)} = 0.068$	$b_{\text{Math}} = 1.32\ (0.13)*$ $b_{\text{gender}} = 11.71\ (1.07)*$ $SD_{\text{lec}} = 6.25$ $SD_{\text{SLR}} = 27.24$ $R_{2\text{GLMM}(c)} = 0.113$	$b_{\text{Math}} = 1.76\ (0.07)*$ $b_{\text{gender}} = 0.43\ (0.53)$ $SD_{\text{lec}} = 3.77$ $SD_{\text{SLR}} = 13.54$ $R_{2\text{GLMM}(c)} = 0.252$	$b_{\text{Math}} = 2.60\ (0.08)*$ $b_{\text{gender}} = -0.97\ (0.62)$ $SD_{\text{lec}} = 1.52$ $SD_{\text{SLR}} = 15.69$ $R_{2\text{GLMM}(c)} = 0.305$	$b_{\text{Math}} = 2.34\ (0.09)*$ $b_{\text{gender}} = 1.72\ (0.73)*$ $SD_{\text{lec}} = 1.38$ $SD_{\text{SLR}} = 18.53$ $R_{2\text{GLMM}(c)} = 0.201$	$b_{\text{Math}} = 2.23\ (0.08)*$ $b_{\text{gender}} = 0.25\ (0.68)$ $SD_{\text{lec}} = 2.90$ $SD_{\text{SLR}} = 17.37$ $R_{2\text{GLMM}(c)} = 0.220$	$b_{\text{Math}} = 1.77\ (0.06)*$ $b_{\text{gender}} = 1.28\ (0.50)*$ $SD_{\text{lec}} = 1.71$ $SD_{\text{SLR}} = 12.64$ $R_{2\text{GLMM}(c)} = 0.241$
URM	$b_{\text{URM}} = -0.59\ (0.65)$ $SD_{\text{lec}} = 0.96$ $SD_{\text{SLR}} = 11.37$ $R_{2\text{GLMM}(c)} = 0.007$	$b_{\text{URM}} = -4.74\ (1.07)*$ $SD_{\text{lec}} = 2.63$ $SD_{\text{SLR}} = 18.84$ $R_{2\text{GLMM}(c)} = 0.026$	$b_{\text{URM}} = -8.75\ (1.65)*$ $SD_{\text{lec}} = 6.23$ $SD_{\text{SLR}} = 28.84$ $R_{2\text{GLMM}(c)} = 0.053$	$b_{\text{URM}} = -6.89\ (0.91)*$ $SD_{\text{lec}} = 4.03$ $SD_{\text{SLR}} = 15.92$ $R_{2\text{GLMM}(c)} = 0.077$	$b_{\text{URM}} = -9.82\ (1.09)***$ $SD_{\text{lec}} = 1.51$ $SD_{\text{SLR}} = 19.18$ $R_{2\text{GLMM}(c)} = 0.033$	$b_{\text{URM}} = -9.03\ (1.22)*$ $SD_{\text{lec}} = 1.47$ $SD_{\text{SLR}} = 21.35$ $R_{2\text{GLMM}(c)} = 0.023$	$b_{\text{URM}} = -9.45\ (1.14)*$ $SD_{\text{lec}} = 3.06$ $SD_{\text{SLR}} = 20.07$ $R_{2\text{GLMM}(c)} = 0.044$	$b_{\text{URM}} = -7.13\ (0.86)*$ $SD_{\text{lec}} = 1.98$ $SD_{\text{SLR}} = 15.05$ $R_{2\text{GLMM}(c)} = 0.039$
Math ACT + URM	$b_{\text{Math}} = 0.23\ (0.05)*$ $b_{\text{URM}} = -0.33\ (0.66)$ $SD_{\text{lec}} = 0.67$ $SD_{\text{SLR}} = 10.78$ $R_{2\text{GLMM}(c)} = 0.011$	$b_{\text{Math}} = 0.80\ (0.09)*$ $b_{\text{URM}} = -2.99\ (1.09)*$ $SD_{\text{lec}} = 2.34$ $SD_{\text{SLR}} = 17.70$ $R_{2\text{GLMM}(c)} = 0.054$	$b_{\text{Math}} = 1.11\ (0.14)*$ $b_{\text{URM}} = -5.82\ (1.70)*$ $SD_{\text{lec}} = 5.96$ $SD_{\text{SLR}} = 27.62$ $R_{2\text{GLMM}(c)} = 0.076$	$b_{\text{Math}} = 1.72\ (0.07)*$ $b_{\text{URM}} = -1.87\ (0.83)*$ $SD_{\text{lec}} = 3.85$ $SD_{\text{SLR}} = 13.50$ $R_{2\text{GLMM}(c)} = 0.257$	$b_{\text{Math}} = 2.59\ (0.08)*$ $b_{\text{URM}} = -1.86\ (0.97)*$ $SD_{\text{lec}} = 1.49$ $SD_{\text{SLR}} = 15.74$ $R_{2\text{GLMM}(c)} = 0.307$	$b_{\text{Math}} = 2.29\ (0.09)*$ $b_{\text{URM}} = -2.42\ (1.15)*$ $SD_{\text{lec}} = 1.45$ $SD_{\text{SLR}} = 18.57$ $R_{2\text{GLMM}(c)} = 0.203$	$b_{\text{Math}} = 2.15\ (0.09)*$ $b_{\text{URM}} = -3.68\ (1.07)*$ $SD_{\text{lec}} = 2.89$ $SD_{\text{SLR}} = 17.30$ $R_{2\text{GLMM}(c)} = 0.225$	$b_{\text{Math}} = 1.71\ (0.06)*$ $b_{\text{URM}} = -2.33\ (0.78)*$ $SD_{\text{lec}} = 1.80$ $SD_{\text{SLR}} = 12.65$ $R_{2\text{GLMM}(c)} = 0.245$
FG	$b_{\text{FG}} = -0.28\ (0.49)$ $SD_{\text{lec}} = 0.94$ $SD_{\text{SLR}} = 11.23$ $R_{2\text{GLMM}(c)} = 0.007$	$b_{\text{FG}} = -2.05\ (0.82)*$ $SD_{\text{lec}} = 2.52$ $SD_{\text{SLR}} = 18.90$ $R_{2\text{GLMM}(c)} = 0.019$	$b_{\text{FG}} = -2.16\ (1.26)$ $SD_{\text{lec}} = 6.31$ $SD_{\text{SLR}} = 29.13$ $R_{2\text{GLMM}(c)} = 0.046$	$b_{\text{FG}} = -3.97\ (0.69)*$ $SD_{\text{lec}} = 3.90$ $SD_{\text{SLR}} = 15.98$ $R_{2\text{GLMM}(c)} = 0.066$	$b_{\text{FG}} = -6.81\ (0.83)*$ $SD_{\text{lec}} = 1.50$ $SD_{\text{SLR}} = 19.15$ $R_{2\text{GLMM}(c)} = 0.028$	$b_{\text{FG}} = -6.09\ (0.92)*$ $SD_{\text{lec}} = 1.44$ $SD_{\text{SLR}} = 21.34$ $R_{2\text{GLMM}(c)} = 0.019$	$b_{\text{FG}} = -6.32\ (0.87)*$ $SD_{\text{lec}} = 3.04$ $SD_{\text{SLR}} = 20.17$ $R_{2\text{GLMM}(c)} = 0.038$	$b_{\text{FG}} = -4.52\ (0.65)*$ $SD_{\text{lec}} = 1.87$ $SD_{\text{SLR}} = 15.06$ $R_{2\text{GLMM}(c)} = 0.030$
Math ACT + FG	$b_{\text{Math}} = 0.23\ (0.05)*$ $b_{\text{FG}} = -0.25\ (0.50)$ $SD_{\text{lec}} = 0.65$ $SD_{\text{SLR}} = 10.67$ $R_{2\text{GLMM}(c)} = 0.011$	$b_{\text{Math}} = 0.82\ (0.09)*$ $b_{\text{FG}} = -1.13\ (0.83)$ $SD_{\text{lec}} = 2.26$ $SD_{\text{SLR}} = 17.73$ $R_{2\text{GLMM}(c)} = 0.050$	$b_{\text{Math}} = 1.19\ (0.14)*$ $b_{\text{FG}} = -1.16\ (1.30)$ $SD_{\text{lec}} = 6.20$ $SD_{\text{SLR}} = 27.83$ $R_{2\text{GLMM}(c)} = 0.074$	$b_{\text{Math}} = 1.73\ (0.07)*$ $b_{\text{FG}} = -1.32\ (0.63)*$ $SD_{\text{lec}} = 3.78$ $SD_{\text{SLR}} = 13.53$ $R_{2\text{GLMM}(c)} = 0.254$	$b_{\text{Math}} = 2.56\ (0.08)*$ $b_{\text{FG}} = -2.67\ (0.73)*$ $SD_{\text{lec}} = 1.49$ $SD_{\text{SLR}} = 15.66$ $R_{2\text{GLMM}(c)} = 0.308$	$b_{\text{Math}} = 2.28\ (0.09)*$ $b_{\text{FG}} = -2.34\ (0.86)*$ $SD_{\text{lec}} = 1.40$ $SD_{\text{SLR}} = 18.52$ $R_{2\text{GLMM}(c)} = 0.202$	$b_{\text{Math}} = 2.17\ (0.09)*$ $b_{\text{FG}} = -2.94\ (0.81)*$ $SD_{\text{lec}} = 2.88$ $SD_{\text{SLR}} = 17.32$ $R_{2\text{GLMM}(c)} = 0.224$	$b_{\text{Math}} = 1.72\ (0.06)*$ $b_{\text{FG}} = -1.89\ (0.59)*$ $SD_{\text{lec}} = 1.74$ $SD_{\text{SLR}} = 12.63$ $R_{2\text{GLMM}(c)} = 0.242$

TABLE XI. TABLE A2. Means, standard deviations, regression coefficients, and conditional R2 values for multiple regression models using demographics and Math ACT scores to predict course grade components in the calculus-based course.

Predictor: Calculus-based course	Lab	HW	ES	Quizzes	Midterm 1	Midterm 2	Final exam	Total Pts
	$M = 95.7$ $SD = 12.9$	$M = 86.4$ $SD = 19.1$	$M = 84.1$ $SD = 24.1$	$M = 76.3$ $SD = 18.3$	$M = 76.2$ $SD = 18.4$	$M = 69.8$ $SD = 22.5$	$M = 65.5$ $SD = 24.0$	$M = 76.8$ $SD = 16.4$
Gender	$b_{\text{gender}} = 2.13$ (0.44)*	$b_{\text{gender}} = 4.01$ (0.68)*	$b_{\text{gender}} = 8.29$ (0.86)*	$b_{\text{gender}} = -2.22$ (0.64)*	$b_{\text{gender}} = -2.45$ (0.66)*	$b_{\text{gender}} = -2.95$ (0.79)*	$b_{\text{gender}} = -1.38$ (0.83)	$b_{\text{gender}} = -0.34$ (0.57)
	$SD_{\text{lec}} = 1.51$	$SD_{\text{lec}} = 1.93$	$SD_{\text{lec}} = 5.86$	$SD_{\text{lec}} = 5.47$	$SD_{\text{lec}} = 3.92$	$SD_{\text{lec}} = 4.78$	$SD_{\text{lec}} = 6.58$	$SD_{\text{lec}} = 2.54$
	$SD_{\text{SLR}} = 11.61$	$SD_{\text{SLR}} = 17.91$	$SD_{\text{SLR}} = 22.55$	$SD_{\text{SLR}} = 16.62$	$SD_{\text{SLR}} = 17.15$	$SD_{\text{SLR}} = 20.54$	$SD_{\text{SLR}} = 21.66$	$SD_{\text{SLR}} = 14.95$
	$R_{2\text{GLMM}(c)} = 0.023$	$R_{2\text{GLMM}(c)} = 0.020$	$R_{2\text{GLMM}(c)} = 0.084$	$R_{2\text{GLMM}(c)} = 0.100$	$R_{2\text{GLMM}(c)} = 0.053$	$R_{2\text{GLMM}(c)} = 0.055$	$R_{2\text{GLMM}(c)} = 0.085$	$R_{2\text{GLMM}(c)} = 0.028$
Math ACT + Gender	$b_{\text{Math}} = 0.22$ (0.06)*	$b_{\text{Math}} = 0.92$ (0.09)*	$b_{\text{Math}} = 0.70$ (0.11)*	$b_{\text{Math}} = 2.05$ (0.08)*	$b_{\text{Math}} = 2.06$ (0.08)*	$b_{\text{Math}} = 2.36$ (0.10)*	$b_{\text{Math}} = 2.59$ (0.02)*	$b_{\text{Math}} = 1.80$ (0.07)*
	$b_{\text{gender}} = 1.92$ (0.44)*	$b_{\text{gender}} = 4.38$ (0.68)*	$b_{\text{gender}} = 8.20$ (0.87)*	$b_{\text{gender}} = -1.20$ (0.58)*	$b_{\text{gender}} = -1.30$ (0.61)*	$b_{\text{gender}} = -1.69$ (0.73)*	$b_{\text{gender}} = -0.14$ (0.76)	$b_{\text{gender}} = 0.52$ (0.52)
	$SD_{\text{lec}} = 1.29$	$SD_{\text{lec}} = 1.85$	$SD_{\text{lec}} = 5.47$	$SD_{\text{lec}} = 5.32$	$SD_{\text{lec}} = 4.00$	$SD_{\text{lec}} = 4.61$	$SD_{\text{lec}} = 6.33$	$SD_{\text{lec}} = 2.27$
	$SD_{\text{SLR}} = 10.95$	$SD_{\text{SLR}} = 16.96$	$SD_{\text{SLR}} = 21.68$	$SD_{\text{SLR}} = 14.42$	$SD_{\text{SLR}} = 15.19$	$SD_{\text{SLR}} = 18.27$	$SD_{\text{SLR}} = 19.00$	$SD_{\text{SLR}} = 12.93$
	$R_{2\text{GLMM}(c)} = 0.022$	$R_{2\text{GLMM}(c)} = 0.050$	$R_{2\text{GLMM}(c)} = 0.089$	$R_{2\text{GLMM}(c)} = 0.265$	$R_{2\text{GLMM}(c)} = 0.215$	$R_{2\text{GLMM}(c)} = 0.200$	$R_{2\text{GLMM}(c)} = 0.238$	$R_{2\text{GLMM}(c)} = 0.194$
URM	$b_{\text{URM}} = -0.70$ (0.67)	$b_{\text{URM}} = -3.98$ (1.03)*	$b_{\text{URM}} = -1.57$ (1.30)	$b_{\text{URM}} = -8.91$ (0.94)*	$b_{\text{URM}} = -8.39$ (0.98)*	$b_{\text{URM}} = -9.23$ (1.17)*	$b_{\text{URM}} = -10.15$ (1.23)*	$b_{\text{URM}} = -7.31$ (0.85)*
	$SD_{\text{lec}} = 1.39$	$SD_{\text{lec}} = 1.77$	$SD_{\text{lec}} = 5.66$	$SD_{\text{lec}} = 5.46$	$SD_{\text{lec}} = 3.72$	$SD_{\text{lec}} = 4.66$	$SD_{\text{lec}} = 6.64$	$SD_{\text{lec}} = 2.47$
	$SD_{\text{SLR}} = 11.69$	$SD_{\text{SLR}} = 17.94$	$SD_{\text{SLR}} = 22.70$	$SD_{\text{SLR}} = 16.41$	$SD_{\text{SLR}} = 16.99$	$SD_{\text{SLR}} = 20.37$	$SD_{\text{SLR}} = 21.42$	$SD_{\text{SLR}} = 14.77$
	$R_{2\text{GLMM}(c)} = 0.014$	$R_{2\text{GLMM}(c)} = 0.014$	$R_{2\text{GLMM}(c)} = 0.059$	$R_{2\text{GLMM}(c)} = 0.119$	$R_{2\text{GLMM}(c)} = 0.064$	$R_{2\text{GLMM}(c)} = 0.065$	$R_{2\text{GLMM}(c)} = 0.103$	$R_{2\text{GLMM}(c)} = 0.046$
Math ACT + URM	$b_{\text{Math}} = 0.19$ (0.06)*	$b_{\text{Math}} = 0.86$ (0.09)*	$b_{\text{Math}} = 0.60$ (0.12)*	$b_{\text{Math}} = 1.98$ (0.08)*	$b_{\text{Math}} = 2.03$ (0.08)*	$b_{\text{Math}} = 2.30$ (0.10)*	$b_{\text{Math}} = 2.54$ (0.10)*	$b_{\text{Math}} = 1.75$ (0.07)*
	$b_{\text{URM}} = -0.43$ (0.69)	$b_{\text{URM}} = -2.15$ (1.07)*	$b_{\text{URM}} = 0.09$ (1.38)	$b_{\text{URM}} = -4.08$ (0.90)*	$b_{\text{URM}} = -3.14$ (0.95)*	$b_{\text{URM}} = -3.77$ (1.15)*	$b_{\text{URM}} = -3.87$ (1.19)*	$b_{\text{URM}} = -3.05$ (0.81)*
	$SD_{\text{lec}} = 1.24$	$SD_{\text{lec}} = 1.78$	$SD_{\text{lec}} = 5.17$	$SD_{\text{lec}} = 5.28$	$SD_{\text{lec}} = 3.88$	$SD_{\text{lec}} = 4.56$	$SD_{\text{lec}} = 6.38$	$SD_{\text{lec}} = 2.26$
	$SD_{\text{SLR}} = 11.09$	$SD_{\text{SLR}} = 17.13$	$SD_{\text{SLR}} = 21.95$	$SD_{\text{SLR}} = 14.43$	$SD_{\text{SLR}} = 15.20$	$SD_{\text{SLR}} = 18.32$	$SD_{\text{SLR}} = 19.01$	$SD_{\text{SLR}} = 12.96$
	$R_{2\text{GLMM}(c)} = 0.016$	$R_{2\text{GLMM}(c)} = 0.041$	$R_{2\text{GLMM}(c)} = 0.060$	$R_{2\text{GLMM}(c)} = 0.268$	$R_{2\text{GLMM}(c)} = 0.218$	$R_{2\text{GLMM}(c)} = 0.201$	$R_{2\text{GLMM}(c)} = 0.245$	$R_{2\text{GLMM}(c)} = 0.201$
FG	$b_{\text{FG}} = -1.36$ (0.47)*	$b_{\text{FG}} = -4.14$ (0.73)*	$b_{\text{FG}} = -2.39$ (0.93)*	$b_{\text{FG}} = -5.30$ (0.76)*	$b_{\text{FG}} = -3.97$ (0.70)*	$b_{\text{FG}} = -6.37$ (0.83)*	$b_{\text{FG}} = -7.26$ (0.88)*	$b_{\text{FG}} = -5.02$ (0.60)*
	$SD_{\text{lec}} = 1.41$	$SD_{\text{lec}} = 1.76$	$SD_{\text{lec}} = 5.68$	$SD_{\text{lec}} = 5.30$	$SD_{\text{lec}} = 3.80$	$SD_{\text{lec}} = 4.76$	$SD_{\text{lec}} = 6.63$	$SD_{\text{lec}} = 2.53$
	$SD_{\text{SLR}} = 11.62$	$SD_{\text{SLR}} = 17.90$	$SD_{\text{SLR}} = 22.89$	$SD_{\text{SLR}} = 16.51$	$SD_{\text{SLR}} = 17.14$	$SD_{\text{SLR}} = 20.42$	$SD_{\text{SLR}} = 21.46$	$SD_{\text{SLR}} = 14.81$
	$R_{2\text{GLMM}(c)} = 0.017$	$R_{2\text{GLMM}(c)} = 0.018$	$R_{2\text{GLMM}(c)} = 0.060$	$R_{2\text{GLMM}(c)} = 0.113$	$R_{2\text{GLMM}(c)} = 0.055$	$R_{2\text{GLMM}(c)} = 0.065$	$R_{2\text{GLMM}(c)} = 0.102$	$R_{2\text{GLMM}(c)} = 0.045$
Math ACT + FG	$b_{\text{Math}} = 0.17$ (0.06)*	$b_{\text{Math}} = 0.81$ (0.09)*	$b_{\text{Math}} = 0.54$ (0.11)*	$b_{\text{Math}} = 2.01$ (0.08)*	$b_{\text{Math}} = 2.07$ (0.08)*	$b_{\text{Math}} = 2.32$ (0.10)*	$b_{\text{Math}} = 2.53$ (0.10)*	$b_{\text{Math}} = 1.75$ (0.07)*
	$b_{\text{FG}} = -0.87$ (0.48)	$b_{\text{FG}} = -2.66$ (0.75)*	$b_{\text{FG}} = -1.59$ (0.97)	$b_{\text{FG}} = -2.00$ (0.64)*	$b_{\text{FG}} = -0.79$ (0.67)	$b_{\text{FG}} = -2.52$ (0.81)*	$b_{\text{FG}} = -2.83$ (0.84)*	$b_{\text{FG}} = -2.06$ (0.60)*
	$SD_{\text{lec}} = 1.17$	$SD_{\text{lec}} = 1.68$	$SD_{\text{lec}} = 5.18$	$SD_{\text{lec}} = 5.33$	$SD_{\text{lec}} = 3.94$	$SD_{\text{lec}} = 4.60$	$SD_{\text{lec}} = 6.35$	$SD_{\text{lec}} = 2.26$
	$SD_{\text{SLR}} = 10.96$	$SD_{\text{SLR}} = 17.02$	$SD_{\text{SLR}} = 21.97$	$SD_{\text{SLR}} = 14.40$	$SD_{\text{SLR}} = 15.20$	$SD_{\text{SLR}} = 18.26$	$SD_{\text{SLR}} = 18.95$	$SD_{\text{SLR}} = 12.90$
	$R_{2\text{GLMM}(c)} = 0.016$	$R_{2\text{GLMM}(c)} = 0.041$	$R_{2\text{GLMM}(c)} = 0.060$	$R_{2\text{GLMM}(c)} = 0.267$	$R_{2\text{GLMM}(c)} = 0.216$	$R_{2\text{GLMM}(c)} = 0.201$	$R_{2\text{GLMM}(c)} = 0.243$	$R_{2\text{GLMM}(c)} = 0.199$

[1] National Center for Science, and Engineering Statistics, *Women, Minorities, and Persons with Disabilities in Science, and Engineering: Special Report NSF 19-340* (National Science Foundation, Washington, DC, 2019).

[2] L. Merner and J. Tyler, *Native American Participation among Bachelors in Physical Sciences and Engineering: Results from 2003-13 Data of the National Center for Education Statistics. Focus On* (AIP Statistical Research Center, New York, 2017).

[3] S. Cheryan, S. A. Ziegler, A. K. Montoya, and L. Jiang, Why are some STEM fields more gender balanced than others?, Psychol. Bull. **143**, 1 (2017).

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

[5] C. Riegle-Crumb, B. King, and Y. Irizarry, Does STEM stand out? Examining racial/ethnic gaps in persistence across postsecondary fields, Educ. Res. **48**, 133 (2019).

[6] E. B. Witherspoon, P. Vincent-Ruz, and C. D. Schunn, When making the grade isn't enough: The gendered nature of premed science course attrition, Educ. Res. **48**, 193 (2019).

[7] G. Stoet and D. C. Geary, The gender-equality paradox in science, technology, engineering, and mathematics education, Psychol. Sci. **29**, 581 (2018).

[8] A. D. Kugler, C. H. Tinsley, and O. Ukhaneva, *Choice of Majors: Are Women Really Different from Men? (No. w23735)* (National Bureau of Economic Research, Cambridge, MA, 2017).

[9] 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).

[10] B. Ost, The role of peers and grades in determining major persistence in the sciences, Econ. Educ. Rev. **29**, 923 (2010).

[11] M. T. Wang, J. S. Eccles, and S. Kenny, Not lack of ability but more choice: Individual and gender differences in choice of careers in science, technology, engineering, and mathematics, Psychol. Sci. **24**, 770 (2013).

[12] R. L. Matz, B. P. Koester, S. Fiorini, G. Grom, L. Shepard, C. G. Stangor, and T. A. McKay, Patterns of gendered performance differences in large introductory courses at five research universities, AERA Open **3**, 1 (2017).

[13] S. Lauer, J. Momsen, E. Offerdahl, M. Kryjevskaia, W. Christensen, and L. Montplaisir, Stereotyped: Investigating gender in introductory science courses, CBE Life Sci. Educ. **12**, 30 (2013).

[14] 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).

[15] 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).

[16] T. G. Greene, C. N. Marti, and K. McClenney, 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).

[17] E. A. Canning, K. Muenks, D. J. Green, and M. C. Murphy, STEM faculty who believe ability is fixed have larger racial achievement gaps and inspire less student motivation in their classes, Sci. Adv. **5**, 4734 (2019).

[18] C. Astorne-Figari and J. D. Speer, Are changes of major major changes? The roles of grades, gender, and preferences in college major switching, Econ. Educ. Rev. **70**, 75 (2019).

[19] T. Ahn, P. Arcidiacono, A. Hopson, and J. R. Thomas, Equilibrium grade inflation with implications for female interest in stem majors, National Bureau of Economic Research working paper 26556 (2019).

[20] B. King, Changing college majors: Does it happen more in STEM and do grades matter?. J. Coll. Sci. Teach. **044**, 44 (2015).

[21] R. Stinebrickner and T. R. Stinebrickner, A major in science? Initial beliefs and final outcomes for college major and dropout, Rev. Econ. Studies **81**, 426 (2014).

[22] K. Rask and J. Tiefenthaler, The role of grade sensitivity in explaining the gender imbalance in undergraduate economics, Econ. Educ. Rev. **27**, 676 (2008).

[23] G. Kortemeyer, Gender differences in the use of an online homework system in an introductory physics course, Phys. Rev. ST Phys. Educ. Res. **5**, 010107 (2009).

[24] S. Salehi, S. Cotner, S. M. Azarin, E. E. Carlson, M. Driessen, V. E. Ferry, and C. J. Ballen, Gender performance gaps across different assessment methods and the underlying mechanisms: The case of incoming preparation and test anxiety, Frontiers Educ. **4**, 107 (2019).

[25] B. D. Mikula and A. F. Heckler, Framework and implementation for improving physics essential skills via computer-based practice: Vector math, Phys. Rev. Phys. Educ. Res. **13**, 010122 (2017).

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

[27] S. L. Dika and M. M. D'Amico, Early experiences and integration in the persistence of first-generation college students in STEM and non-STEM majors, J. Res. Sci. Teach. **53**, 368 (2016).

[28] X. Chen, Students Who Study Science, Technology, Engineering, and Mathematics (STEM) in Postsecondary Education, Stats in Brief. NCES 2009–161 (National Center for Education Statistics, Washington, DC, 2009).

[29] S. Nakagawa and H. Schielzeth, A general and simple method for obtaining R2 from generalized linear mixed-effects models, Methods Ecol. Evol. **4**, 133 (2013).

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

[31] C. J. Ballen, S. Salehi, and S. Cotner, Exams disadvantage women in introductory biology, PLoS One **12**, e0186419 (2017).

[32] S. Cotner and C. J. Ballen, Can mixed assessment methods make biology classes more equitable?, PLoS One **12**, e0189610 (2017).

[33] Test scores do not equal merit: executive summary (2007, August 22). Retrieved from http://fairtest.org/test-scores-do-not-equal-merit-executive-summary.

[34] L. McNeil and P. Heron, Preparing physics students for 21st-century careers, Phys. Today **70**, 38 (2017).