High-stakes exams inflate a gender gap and contribute to systematic grading errors in one introductory physics series

David J. Webb
Department of Physics and Astronomy, University of California-Davis, Davis, California 95616, USA

Cassandra A. Paul
Physics and Astronomy Department, Science Education Program, San Jose State University, San José, California 95192-0106, USA

(Received 15 April 2025; accepted 3 October 2025; published 21 October 2025)

Previous research has suggested that changing the percentage of the course grade associated with exam grades in science, technology, engineering, and mathematics courses can change the gender gap in the course. It has also been shown that high-stakes assessments yield the lowest (relative) scores for female students. Previous research by the authors has shown that implementing retake exams can eliminate the gender gap in introductory physics courses. This paper explores several different hypotheses for why retake exams are associated with a zeroed gender gap. Two independent measurements comparing exams with different stakes are used in support of the argument that the entire gender gap on introductory physics exams may be due to the stakes associated with those exams. In other words, these data support the idea that a gender grade gap on exams does not measure a gender difference in the physics knowledge or physics ability of these students. Implications suggest that instructors should choose lower stakes assessment options if they are interested in exam measurements that are not influenced by differences in students' performance related to exam stakes.

DOI: 10.1103/ngrs-dxtf

I. INTRODUCTION

In a recent paper [1], we showed data in support of the argument that demographic grade gaps are a result of course design (the course deficit model [2]) rather than the more common argument that the gaps result from average deficits in specific demographic groups of students (the student deficit model [3]). One way we did this was by showing that the gender gap disappeared in introductory physics classes that were changed to offer exam retake opportunities to all students, while the gender gap had been persistent in regular, unchanged courses. Our previous work did not examine the mechanism for how the retakes eliminated the gender gap. For example, were women more likely to take advantage of the retake opportunity, or did they show greater improvement than men on the second exam? In this paper, we analyze more detailed assessment data on the gender gap issue in these courses and argue that a reasonable interpretation of our results is that the entire gender gap in these courses is determined by the stakes associated with the exams and does not reflect a gender difference in physics knowledge, opportunity, or ability.

A. Gender gaps in physics

Gender equity gaps in introductory physics courses and exams have been studied extensively over the past several decades. Gender gaps in physics can refer to differences in course grades [1,4], performance on concept inventories [5,6], and/or exam scores [7]. Representation and retention gaps have also been observed [8].

1. Gaps—student-level characteristics

Many different explanations for the existence of these types of gender gaps have been proposed and investigated. Some of the leading theories for the gender gaps in physics are that women on average (i) have poorer physics and/or math preparation [6,9,10], (ii) have less overall interest in physics [11], and/or (iii) suffer more from test anxiety, which is amplified on high-stakes exams [7,10,12–16]. It has also been shown that women report a lower sense of belonging in physics compared to their male peers [4,17], a weaker science or physics identity [18], and/or lower self-efficacy in physics [10,15,19,20]. Finally, the context (e.g., physical vs biological or practical vs theoretical) of an exam question may influence [21,22] the gender gap for that question.

In a comprehensive review of gender gaps in concept inventories, Madsen *et al.* [5] found that none of the 30 different factors they considered fully explained the gender gap and thus concluded that there may be many interconnected issues leading to these gaps. Despite these different competing factors, there have been a few studies that have seemed to entirely eliminate the gender gap [1], suggesting that, despite the gap's persistence and the uncertainty concerning its cause, it is possible to mitigate the effects giving rise to the gap.

Exploring one potentially fruitful mitigation of gender gaps in exams, Cotner and Ballen [2] show that exam gender gaps in some biological science courses could be changed by lowering the stakes associated with exams. They compared overall success in different biology courses by categorizing them as either low-stakes (less than 40% of the course grade determined by quizzes and exams) or high stakes (more than 40% of the course grade determined by quizzes and exams). The men outperformed the women in the high-stakes courses, while there was no statistical difference between the genders in low-stakes courses.

Cotner *et al.* later suggest [12] that the reason for the change in performance when the stakes change is because women are more likely to suffer from anxiety, which is amplified in high-stakes situations. In fact, the connection between high-stakes exams and anxiety in women is well studied [7,10,12–16], where evidence suggests that women are more likely to suffer from test anxiety that negatively impacts exam performance [23]. Malespina and Singh [10] found that women underperformed on high-stakes exams in calculus-based introductory physics courses. However, when they controlled for test anxiety, the gap became nonsignificant. There is evidence that test anxiety is an issue for women in other science, technology, engineering, and mathematics (STEM) majors as well. Espinosa *et al.* [15] found that in a large introductory biology course, women expressed higher test anxiety than men, and that anxiety was a significant predictor of exam grades even when controlling for overall grade point average (GPA). This is important not only in biology courses but also for introductory physics courses, as a large population of introductory physics students come from biology-related majors. Similarly, Malespina *et al.* [14] found the same pattern among bioscience students enrolled in a physics course. Salehi *et al.* [13] examined performance across three different assessment types—exams, nonexams, and labs—in five different lower division STEM courses. They found that while there was a significant negative gap for women on exams, women consistently outperformed men on the other two types of assessments. Notably, they found that exam performance for women was negatively influenced by test anxiety in both lower and upper division courses. This was true even in courses where the gender gap in exam scores had disappeared, including both biology and other STEM subjects.

2. *Gaps—system level issues*

Leslie *et al.* [24] take a more systemic view in noting that gender gaps in representation are largest in a few specific academic fields (including physics) where the practitioners in the field are more likely to think that to succeed in their field "hard work alone just won't cut it; you need to have an innate gift or talent." Muradoglu *et al.* [25] use this idea, that success in a field is perceived to require innate brilliance, together with the general societal view that women are not as intellectually talented as men [26], to argue that a natural result is that, on average, women would have lower average self-efficacy in such fields. Since self-efficacy inversely correlates with anxiety, it is also reasonable that women, on average, would be expected to be more anxious than men during the high-stakes exams in these particular academic fields. As noted earlier, in this paper, we will show data consistent with the idea that there is no gender gap in physics understanding or ability, but that the commonly measured exam gender gap arises because of the stakes associated with that exam. In this sense, the exam gender gap may simply be an incorrect measurement of the underlying knowledge or ability. It would be ironic if such illusory exam gender grade gaps were part of the evidence society used to conclude that women were not as talented as men at physics and that conclusion led to women being more anxious on, and underperforming on, exactly these high-stakes exams that led to society's conclusions.

3. *Changes in course structure changes gaps*

Although the evidence for anxiety may make it a compelling explanation for the gender gap, we will continue our previous argument [1] that it is the course structure that determines demographic grade gaps. We suggest that high-stakes exams be considered the cause of the gender gap. One reason we take this position is that high-stakes exams (being characteristics of the course structure) are easily changed, while anxiety (being a characteristic of the student) is much more difficult to change. If high-stakes exams are the cause of the gender gap, changes in course structure can serve as a mitigation.

As an example of why we take this approach, consider a study performed by Simmons and Heckler [27], where they simulated new course grades for students (after a course had already been completed) by reducing the weight of the exams in the overall course grade. They found that making the exams worth less of the overall course grade, equity gaps in the course were reduced for women and other historically marginalized students compared to their nonmarginalized peers. In this example, women's anxiety was not changed, but instead the course structure was changed to successfully address the inequity. While this may seem like a semantics argument, positioning the stakes as the possible cause of the gender gap (instead of anxiety) focuses the research and intervention efforts on the changeable aspects of the course, rather than the (potentially unchangeable) aspects of entire populations. Similarly, it removes the urge to “fix” any perceived inadequacies among an underrepresented group (in this case, women).

B. Retake exams in physics

If high stakes are the cause of the gender gap, one obvious solution would be to eliminate the exams as an assessment practice. Critics of this approach say that low-stakes assessments (like homework or group projects) are not as accurate as other forms of assessment because low-stakes assessments like homework allow students to use outside resources, collaboratively solve problems, and potentially even find solutions they can copy without understanding. Others claim that there is value in measuring what a student can do without outside resources, and/or under a high-pressure environment—something that homework assignments typically do not measure. While many, including Ref. [23] and also the authors of this paper, disagree with these critiques, if exams (timed solo tests or quizzes with little or no resources allowed) are measuring something important to many instructors, eliminating them entirely may not be a viable option. If so, mitigating the harms or biases caused by these exams may be the most practical approach.

One possible way to lower the stakes of an exam—while still retaining exams as a form of assessment—is to offer a retake option for each exam given during the term. For each in-term exam, a retake covers the same topics as the first-attempt exam, is offered after the students have received their first-attempt scores, is offered to every student regardless of grade, and the retake score can supplant the first-attempt score. Thus far, there is very little research on retake exams in large introductory STEM classes. Davis *et al.* [28] find that students in a large introductory human physiology course who took advantage of retake exams on average increase their grade by 6%. Kortemeyer *et al.* [29] share the benefits of retakes as learning tools in introductory physics. However, none of the examples we found examined the relationship between the retakes and the gender gap in physics courses aside from our own previous work on this same dataset.

C. Models of achievement and assessment equity

This work, like our prior research [1] utilizes two theoretical models. We define a course as being equitable if student grades are not predictable by demographic, meaning that all demographic groups achieve the same overall grades on average. This is the definition of an equity of parity model of equity [30]. We also use Cotner and Ballen’s [2] course deficit model. A course deficit model assumes that any gaps in achievement within the course between any demographic groups are a product of the course structure. Importantly, this does not mean that inequities prior to the course do not exist; the course deficit model merely assumes that the equity gaps measured within the course are a product of the course itself. Our previous work [1] shows that by changing course structures, it is possible to eliminate equity gaps and thus achieve equity of parity without changing the course content or the intellectual level of the content.

We have previously studied grading and assessment in this same physics course and have found that percent scale grading provides an additional penalty to underrepresented minority groups when compared to 4.0 scale grading [31,32]. We also find [33] that on average, women leave slightly more blanks on exams than men, while men are more likely to miss entire quizzes. Given that there seem to be both structural biases and behavioral test-taking differences between different demographic groups, it seems prudent to examine the impact of certain assessment strategies on different demographics.

II. RESEARCH QUESTIONS

In this paper, we make the assumption that because the gender gap decreased or was entirely eliminated in courses that use a retake exam option, the reason for increased gender gaps in the introductory physics courses in our dataset is likely due to the assessment strategy in courses. This leads us to a few different research questions:

1. Does the gender gap decrease, in course sections that allow retakes on exams, because women students are more likely to take advantage of the retake opportunity?
2. Does the gender gap decrease, in course sections that allow retakes on exams, because women students are doing better on their second try?
3. How are the stakes of an exam correlated with the gender gap?

Each research question tests an assumption that either came from the literature or was proposed to the authors of this paper in conversations about our previous work [1]. The first research question examines the assumption that women and men have behavioral differences that influence how likely they are to take advantage of optional opportunities to improve their grades. The second research question tests the assumption that because women are less likely to be prepared to take physics, they might need more time to practice and receive help before succeeding than men do. The third research question was born out of negative results from the first two questions.

III. DATABASE DETAILS

All of the exam data we discuss here are from classes in the introductory physics series taken mostly by bioscience students and offered between 1997 and 2015 at an R1 public university. These dates were chosen because these years included the retake classes, and because the classes that allowed students to take retake exams are part of the same course series as those that did not allow retakes. The topics covered in the course series did not change over the set of years included in the study, and most of the specific student activities are also essentially the same over those years. We chose this set of years because we had access to individual exam scores for almost all of those years, but for no later years. We have access to individual exam scores (both first-take and retake scores) for three of the four classes that offered students retake exams. In our previous paper [1], we argued that the retake effect appeared to be independent of the instructor because these instructors also taught nonretake classes and their course grades showed clear gender gaps in those classes. Finally, exams in these courses are open-ended in the sense that the students are expected to provide both a written explanation of a given problem and perform relevant calculations. All of the exams given in the retake courses were drawn from a pool of exams that had been given previously in these courses, so we consider them to be representative of the usual exams in these classes.

To compare exam scores from retake classes with those from regular classes that did not offer retakes, we added to the database exam data from 254 nonretake classes in the same course series, for which we have in-term exam scores. We also added each student's gender and race or ethnicity as well as each student's university GPA upon entering the class. These data were obtained from the university administration and included before anonymizing the entire dataset several years ago. Since the numerical exam scores are given in a wide variety of ways by the many instructors of these classes, we have z-scored each exam in each course to give each exam grade distribution the same average, 0, and the same standard deviation, 1.

Altogether the database included 26 783 students with 16 493 identified as female students, 10 254 identified as male students, 5 identified as nonbinary, and 31 students with unknown gender identities. This university started collecting self-identified gender information in the Fall term of 2015, so our dataset is almost completely binary gender only. For completeness, the data regarding the small numbers of unknown gender and nonbinary students are discussed in Appendix F. The exam data include 5598 students who appeared only once, 7382 students who appeared in only two of the classes and only once in each, and 13 268 students who appeared exactly once in each of the three classes in the three-course series. The three retake classes had a total of 397 students.

The number of individual exams given during the term varied from a minimum of 2 midterms given in each of seven classes to a maximum of 9 weekly exams given in each of 6 classes, with the most common number given being 4 in-term exams in each of 115 classes (3 of those are the retake classes). In addition, the databases from 203 classes also had final exam grades, and so those classes had all exams accounted for.

A. Classifying the stakes of an assessment

For the purposes of this paper, we will classify the stakes of an exam into three different categories. We use the same definition of high-stakes assessments as those typically used in the literature [2]. "High-stakes" assessments are timed exams that are usually in-person and allow limited use of resources. For example, a student might be able to bring a note card, but they would not be able to use the Internet during the exam. In addition, these exams are considered high stakes because their scores are a large factor in determining the course grade. This is an implied distinction in previous work because high-stakes exams are often differentiated from quizzes that are also usually timed and allow limited resources [2] but have scores that are not as important factors in the course grade.

Instead of examining the stakes of different assessment types (e.g., homework vs exams) as others have done, in this paper, we are investigating whether the differences in the stakes of different exams have an impact on equity gaps found in those exams. To that end, we will take the stakes of any particular exam to be a measure of the importance of that exam's (or quiz's) score to the course grade. For instance, in the classes included in our database, the grade on the final exam constitutes roughly 50% of the course grade (no courses in our dataset allowed retakes on final exams). Therefore, the final exams in either type of course are considered "high stakes" as compared to other exams.

We also introduce the idea of a medium-stakes assessment. In nonretake courses, the entire set of two to nine in-term exams (in-term meaning excluding the final) constitutes roughly the other 50% of the course grade, so each of those exams has considerably lower stakes than the lone final exam. We consider these exams given as the term progresses to be "medium stakes" both because each individual in-term exam counts for a much smaller portion of the course grade than the final exam and because usually the class syllabus describes in-course grading rules telling students that the lowest in-term exam would be dropped before calculating the course grade.

On the other hand, in retake courses, the entire set of four first-attempt in-term exam scores *may have no contribution at all to the course grade* because the retake exam score is always available to supplant the first-attempt exam score. For this reason, we consider a first-attempt exam in a retake course to have relatively lower stakes than any in-term exams in a nonretake class and therefore define the first attempt on a retake as a "low-stakes" assessment.

We consider the first try on an in-term exam that has a retake option to be a low-stakes assessment because a

TABLE I. Descriptions of the magnitude of the stakes associated with an exam. The stakes associated with one exam can be thought of as determined by the minimum amount that that set of exams can possibly contribute to the course grade, divided by the total number of exams of that type. Retake courses and regular courses each had a single final exam, so all final exams had high stakes.

Exam stakes	Exam type	Minimum possible contribution to course grade by an average exam
Low	**Retake class—**Set of four first attempts at short in-term exams	0%
Medium	**Non-retake class—**Set of four to nine short in-term exams or two midterm exams **Retake class—**Second attempt at in-term exams (i.e., the retakes)	5%–15%
High	**Both types of class—** One long final exam	40%

second try is free to any student and could completely erase that first score. However, we consider the second attempt in a retake class to be a medium-stakes assessment because this grade will contribute to the final grade at roughly the same weight as the in-term exams in the courses that do not have retake exams. All of these definitions of the stakes of different exams are outlined in Table I.

Note that the medium-stakes exams in the retake classes (the actual retake exams) are only taken by a subset of students who choose to take advantage of the opportunity to do so. This will make it difficult to draw conclusions from this dataset, so we will leave any calculations of gender gaps within these student subsets, varying through the term, to Appendix E.

IV. RESULTS

A. In-term exams in retake classes

In a previous paper [1], we showed that the gender gap in course grades was zero for a set of four classes that offered retake exams to their students. In a retake class, students took an in-term exam, got their scores on that exam, and then (the next week) had a chance to take a different exam (the retake exam) that covered the same material. Students were told that if they scored higher on the retake, then the retake score would supplant the grade on their first attempt, but if they scored lower on the retake, then the retake and the first-attempt scores would be averaged to give the recorded exam score. In these classes, there were a total of four in-term exams and four in-term retake exams. Here we take a closer look at the details of these first attempts and the exam scores after the subsequent retakes.

Our expectation was that students who had high scores on their first attempt would not risk lowering that score and so would not take a retake exam. We found that on average, a first attempt of an exam was followed by 34% of the students taking the retake a week later. The retake classes were anomalous in having zero gender gap in course grades (i.e., female students had the same average course grades as male students unlike in the regular courses) suggesting the following possible gap-closing scenario regarding retakes: (i) female students may be more likely to take retake exams to improve their grades and/or (ii) female students may make larger gains on retake exams. Examination of the data shows that the first possibility is wrong. We find that $78\% \pm 4\%$ of male students retook at least one exam and $74\% \pm 3\%$ of female students retook at least one exam (a χ^2 test showed $\chi^2(1, N = 399) = 0.89$ with $p = 0.35$, so no measurable gender difference). All together male students retook an average of 1.41 ± 0.09 exams and female students retook an average of 1.33 ± 0.07 exams (a t test gives $p = 0.49$ so no difference here either). It seems that the two genders had about the same need for retake exams.

We can examine the second possibility mentioned above by measuring the gender gap on each exam and also the gender gap after each retake has been scored. As noted in Sec. III, the grades from the 12 exams from the three different courses were normalized so that (i) each set of first-attempt scores from each exam in each course has the same average of 0 and standard deviation of 1, and (ii) the recorded grades (i.e., after retake) from each exam from each class were similarly z-scored. The result of this is that each gender gap will be measured in units of standard deviations of the exam scores for that exam. We will also control for each student's incoming GPA in our estimates of each in-term exam gender gap. We do this because the exam-score gender gap in a class has a small but clear dependence on the GPA gender gap in that class (see Appendix A which shows these class-dependent gender gap data), and we wish to remove that dependence from our in-term exam results. We control for GPA using hierarchical linear modeling (HLM) with the students as lowest level in the hierarchy, the set of classes as the next higher level, and then the part of the three-course series (A, B, or C) makes up the highest level of the hierarchy. We use all three levels because our retake classes cover all three levels. With HLM, the data are fit first class by class, and then those fits are combined into A, B, and C results, which are then combined to give the final coefficients and their errors. We use HLM to account for class-to-class differences in exams and A, B, and C differences in topics covered, since any of these may be affecting the gender gaps, and we want the final error estimates to reflect all of these variations. For more detailed information on HLM, see Ref. [34]. We use

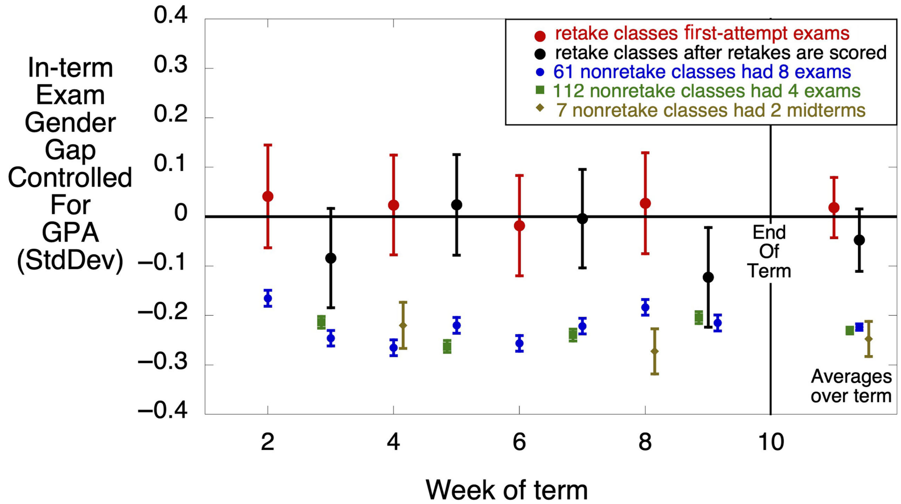

FIG. 1. Each in-term exam was normalized on a class-by-class basis so that the standard deviation of the grade distribution is 1. The gender gap for each group of classes was determined with hierarchical linear modeling while controlling for each student's incoming GPA. This in-term exam score gender gap is plotted as a function of time through the 10-week term at the times the exams and retakes were taken. On the right, after the end of the term, are the overall averages for those first-attempt exams and for the recorded scores after retakes were completed. One sees that the female students generally had higher exam scores than male students (i.e., a positive gender gap) after the first attempt, and that, overall, retakes favored the male students. For reference, the gender gaps for 180 nonretake classes are also shown on a week-by-week basis through their terms.

the variables GPA for a student's GPA and $Female = 1$ for female students and 0 for male students, and fit the following model 8 times, once for each of the four first-attempt exam scores taken through the term and once for each of the four exam scores recorded after the retake scores are factored in,

$$InTrmExm = b_0 + b_{\mathrm{GPA}} GPA + b_{Female} Female \quad (1)$$

so b_{Female} is the (GPA-controlled) gender gap for each fit of in-term exam scores. We also average each student's first-attempt exam scores over the term and fit our model to those, and we average each student's after-retake scores and fit our model to those. So $InTrmExm$ can be one of the four first-attempt exam results, one of the four after-retake exam results or one of the averages over the full term.

Figure 1 shows these gender gaps for each first attempt and also after each retake exam. These results are plotted as a function of the week the exam was taken through the 10-week term. The overall averages, after all four in-term first attempts and their retakes were graded, are shown to the right of the end of the term. Notice that *female students outscored male students on three of the four first-attempt in-term exams*. The average scores across all first-attempt exams is slightly positive, indicating that female students slightly outperformed male students. However, the average score across all in-term exams after retakes is slightly negative, indicating that retake exam benefited male students slightly more than female students. So the idea that the retakes allowed female students to improve their first-attempt scores in a way that reduced the gender gap is also untrue. This result suggests a third possible effect of allowing retake exams. (iii) The knowledge that retakes are allowed lowers the stakes of each first attempt compared to a setting without retakes. So it may be that the scores are gender neutral because the stakes of the first-attempt exams are low. Specifically, we suggest that the first attempt has lower stakes in a retake class than it does in the regular nonretake classes, where the first attempt is also the last attempt. In Fig. 1, we have also included the week-by-week GPA-controlled in-term exam gender gaps for three sets of nonretake classes for which the week-by-week exam pattern through the term is known. In the rest of the paper, we use the ideas of low-stakes, medium-stakes, and high-stakes exams in comparing the retake classes with the 254 nonretake classes for which we have exam data.

B. Low-stakes versus medium-stakes assessments

Instructors in this series of classes determined course grades almost completely by averaging the exam grades using a weighting of their choice. A weighting that was common (almost universal) was to (1) drop the lowest score among the in-term exams (either one in-term exam or one-half of one midterm) and then (2) weigh the final exam somewhere between 40% and 60% depending on whether the in-term exam average was higher or lower, respectively, than the final exam score. The most common numbers of in-term exams were 4 and 8, so the grade weight of a single in-term exam ranged from roughly 5% to 15% of the course grade. We consider these exams to be medium-stakes assessments. The final exam carries several times the weight of any particular in-term exam, so we consider it to be a high-stakes assessment. For the retake courses, we consider the first attempt of every exam set that students take to have low stakes because the student knows that they are free to take another exam and discount the first one if they choose to. In all classes (retake and nonretake), it was not possible to retake the final exam.

We are interested in comparing the gender gap resulting from low-stakes in-term exams with that from medium-stakes in-term exams. For this reason, we compare the gender gap in the *first-attempt exams* from the retake classes with the gender gaps in the 254 nonretake classes in our dataset. Since different classes had different numbers of in-term exams, we will just average the z-scored in-term exam scores for the entire term for each student in each class and model the term-averaged in-term exam z score for each student. Also, remember that the gender gap for in-term exams in a class has a small but clear dependence on the gender gap for GPAs in that class, so we will control for GPA in our analyses.

We model the comparison of the 3 retake classes with the 254 regular classes using the same kind of 3-level hierarchical regression analysis that was previously discussed. Again, the student level is the lowest level, then students are grouped in classes (second level), and the classes are grouped within the course series (A, B, and C in the series identify the highest level). We use all three levels because our database covers all three levels and also because the different classes may have slightly different grade penalties for female students. We again use the variables GPA for a student's GPA, $Female = 1$ for female students and 0 for male students, and now include $Retake = 1$ for the three retake classes and 0 for the 254 regular classes. The model we fit for the student's in-term exam z score is

$$
\begin{aligned}
InTrmExm = b_0 + b_{\mathrm{GPA}} GPA \\
+ b_{Female} Female + b_{Retake} Retake \\
+ b_{Female*Retake} Female * Retake
\end{aligned}
\tag{2}
$$

where the coefficient b_{Female} represents the GPA-controlled gender gap and the coefficient $b_{Female*Retake}$ measures the change in the gender gap when retakes are possible.

TABLE II. The coefficients from an HLM fit to Eq. (2) are shown along with their standard errors, z statistics, and p values. Included are 257 classes that had a total of 57 806 students whose GPAs are known. The interaction term suggests that the gender gap is significantly different (reduced) for the retake classes.

Coefficients	Value	Error	z statistic	p value
b_{GPA}	0.8121	0.0047	173.0	$<10^{-3}$
b_{Female}	−0.2262	0.0046	−49.0	$<10^{-3}$
b_{Retake}	−0.169	0.052	−3.24	0.001
$b_{Female*Retake}$	0.234	0.057	4.10	$<10^{-3}$
b_0	−2.310	0.018	−128.1	$<10^{-3}$

The coefficients resulting from this HLM fit are given in Table II. Recall that we consider the first-attempt exams we are using for the retake classes to be low stakes. In the 254 nonretake classes which had medium-stakes in-term exams, the average grade penalty (b_{Female}) that female students receive over the term is about 0.23 standard deviations (of the exam average distribution). The interaction coefficient ($b_{Female*Retake}$) for classes with low-stakes exams also has magnitude of 0.23 but is positive, indicating that the low-stakes exams net gender gap is, as expected, approximately zero. Seeing a gender gap disappear when the only change seems to be the stakes of the exams might cause one to pause before claiming that the gender gaps in the medium-stakes exams are actually measuring a gap in physics knowledge or ability.

Note that even though the coefficient for $Retake$ is negative, indicating (perhaps) that men do worse under retake conditions, this is only because the grades are z-scored. In our previous paper [1], we showed that, on average, both genders had higher grades in the retake classes. Also, as shown in Appendix C, the gender gap decreases in size in retake classes for each racial or ethnic group sufficiently represented in our data. Finally, as discussed in Appendix B, to determine whether these coefficients or our resulting conclusions are affected by the fact that many students appear multiple times in our database (because they took more than one class in this three-term series), we tested the effects of limiting each student to a single appearance in the database. We find that the coefficients and our resulting conclusions are unchanged.

C. Testing the stakes hypothesis using the highest stakes assessments

Since 203 of the classes in our database have final exam scores, it seems useful to examine those data because the final exam scores contribute much more to the course grade than any in-term exam and so can be considered to be high-stakes exams. Unfortunately, we do not have a low-stakes version of the final exam to compare with the high-stakes versions because even in the courses with retake options on the exams, it was not possible to retake the final exam. However, we do have the medium-stakes in-term exam scores in each class which should be good predictors of final exam scores in that class because they are written and graded by the same instructors and largely cover the same material. In fact, we find a correlation coefficient between the in-term exam average and the final exam score of 0.65 (with $p < 10^{-3}$) but a smaller correlation between student GPA and either in-term or final exam score (both correlation coefficients are about 0.56, $p < 10^{-3}$). We have already suggested that the medium-stakes in-term exam scores include a bias against female students, so it may be that using these scores as predictors of final exam scores will not only control for the demonstrable knowledge or ability of each student in the particular course but may also

TABLE III. The coefficients from an HLM fit to final exam z scores [Eq. (3)] are shown along with their standard errors, z statistics, and p values. Included are 203 classes that had a total of 48 428 students. The interaction term suggests that the gender gap is not significantly different for the retake classes.

Coefficients	Value	Error	z statistic	p value
$b_{InTrmExm}$	0.9656	0.0052	187.4	$<10^{-3}$
b_{Female}	−0.1155	0.0071	−16.2	$<10^{-3}$
b_{Retake}	0.025	0.066	0.38	0.705
$b_{Female*Retake}$	−0.034	0.080	−0.42	0.676
b_0	0.0744	0.0056	13.23	$<10^{-3}$

control for any gender bias inherent in exams. If this were the case, then the controlled final exam gender gap should be approximately zero. To test this idea, we will use HLM to fit the following model:

$$FinalExam = b_0 + b_{InTrmExm}InTrmExm + b_{Female}Female + b_{Retake}Retake + b_{Female*Retake}Female * Retake \quad (3)$$

where we include the variable $Retake$ and the interaction term even though the retake courses had the same sort of high-stakes final exam as all the other courses. If the gender bias on in-term exams is approximately the same as the gender bias on final exams, then the coefficient b_{Female} from this fit should be nearly zero.

The resulting coefficients from the HLM fit to Eq. (3) are given in Table III. The most notable coefficient is b_{Female}, and the most notable thing about this coefficient is that it is quite clearly not zero. In fact, the (in-term exam controlled) final exam gender gap is half the size of the gender gap for the in-term exams alone. This tells us that the full gender gap on the high-stakes final exam is about 50% larger than that of the medium-stakes in-term exams. So the high-stakes final exam has a much larger gender gap than the medium-stakes in-term exams which, themselves, have a much larger gender gap than the low-stakes first-attempt in-term exams in the retake classes. And $b_{Female*Retake}$ is consistent with there being a similar (in-term-exam controlled) gender gap in the high-stakes final exam in the retake classes. Again, it seems likely that the importance (or weight) of the exam (the exam stakes) is closely related to the size of the grade penalty received by female students. Finally, as noted in Appendix B, we have checked that these coefficients and the resulting conclusions hold when we (randomly) include each student only once in our analyses. Again, as in Sec. IV B, it seems that exams with higher stakes have higher gender gaps.

V. DISCUSSION

In our introduction, we present many student-level variables that may contribute to the gender gap in introductory physics. But these factors—preparation, interest, anxiety, stereotype threat, self-efficacy, identity, and belonging—are all deeply intertwined with both broader societal influences and the cultural context of the physics discipline. For instance, results from Leslie *et al.* [24] on the gender imbalance in academic fields (including physics) that are felt to require "innate talent" rather than persistent effort has led these researchers as well as others (Muradoglu *et al.* [25]) to argue that, in fields perceived to require innate talent, women (and other groups that are often intellectually marginalized) have more negative experiences, feel less welcome, have more "imposter" feelings, and are more vulnerable to stereotype threat [35]. This general idea can connect the various gender gap explanations that we have listed above. We postulate that many of these feelings of otherness are exaggerated in a high-pressure environment. With this stated assumption, we can then argue that the reason the exam gender gap exists is not simply because of negative biases against women, but because these biases exist, and students are placed in a high-pressure environment.

Furthermore, an important takeaway from this logic is that exams may not measuring what we assume they are measuring because if they were, performance gaps between demographic groups would not depend on the stakes of the particular exams, as shown in Secs. IV B and IV C. These two sections provide two independent results, each supporting the idea that increasing the stakes of an exam increases the gender gap in a way that is disconnected from differences in physics knowledge or skills. The implications are that if we care about measuring physics knowledge and skills, we need to be looking at ways to reduce or eliminate the stakes associated with our assessments. Since the stakes of a given timed in-person assessment correlate with the size of the gender gap, exams are currently not good measures of physics knowledge and skills across demographic groups.

Another problem with putting one's focus on the individual student-level issues that large groups of students may be having in physics is that this leads naturally to a consideration of those groups of students as deficient or inadequate in some way. Researchers tend to use a student deficit model rather than the idea that these students are just reacting, on average, the way anyone would if they were part of a group stereotyped by society. Indeed, results from Salehi *et al.* [13] support this interpretation—they find that the gender gaps for exams are much larger in general science and engineering courses, where women are less represented, compared to biological science courses, where women are more represented and may therefore feel more like they belong in that space.

Cotner and Ballen [2] cut the Gordian knot made up of interconnected possible student deficits by suggesting the easiest thing to do is to simply change the structure of the course. They proposed using a course deficit model rather than a student deficit model. Our recent work [1,36] supports the idea that changing the course structure is enough to eliminate equity gaps. While studies that examine the effects that society—and, more specifically, disciplinary cultures—on different demographic groups are important and necessary, we should not present these effects as the sole reasons for performance differences between demographic groups. If we move in that direction, we miss the forest for the trees. Perhaps the most important takeaway from our paper is that the solution to decreasing the gender gap in physics exams is not to make women feel less anxious in a stressful environment, it is to change the environment to be less stressful. This is the power that we have as instructors. We might not be able to change the impact society has on women in physics, but we can mitigate the effects of those impacts in our courses. The implications of this work indicate that instructors should strongly consider adopting assessments such as retakes that lower the stakes associated with testing in introductory physics courses. Timed in-person exams are typically anxiety inducing experiences; however, our work shows that if we change the stakes of these exams by allowing a retake, women perform as well as men. This suggests that the stakes are influencing the outcome of exams so that the exams are inaccurate measures of differences between men and women. Instructors who want measurements of physics understanding less influenced by stakes (or anxiety) should adopt assessments with lower stakes.

Our discovery that gender gaps are dependent on the level of stakes associated with timed exams has an extremely important implication for education researchers. It is common for researchers to use prior exam results (e.g., quantitative SAT, ACT, and Advanced Placement exams) as a proxy for preparation for physics. Women's scores on these exams tend to be lower than those of men [37] (This was part of the impetus for proposing research question 2.). Therefore, a common resulting finding from those studies is that women are somehow less prepared for physics than their male peers and that this lack of preparation is what drives the gender gap. An alternative interpretation of the same numerical results is that women are experiencing the same issue in both contexts, and that high-stakes exams are causing the gender gap rather than simply measuring it. Therefore, this study adds evidence to our previous work indicating that it may not make sense to use exam scores as a proxy for preparation when comparing across group demographics [1]. We suggest that future research using exam scores as a proxy for preparation proceed with thoughtful caution.

VI. LIMITATIONS

All of the retake exam classes (as well as the nonretake classes) were offered within the CLASP (collaborative learning through active sensemaking in physics) series of courses [36,38]. These courses are studio type courses in which students spend almost 80% of their class time working in discussion and lab sections, generally in groups of five. The effects of allowing retake exams may well be different in classes that are predominantly lecture-based. In addition, almost the entire course grade in the courses studied was determined by the exam grades in these courses. It could be that in a course in which most of the grade is determined by timed in-person solo assessments, differences of the impact of the stakes of each of those assessments are easier to identify. The data collected in this study are at least a decade old. Since that time, there have been significant changes to the educational landscape including advances in AI, a worldwide pandemic, and a changing student population. These and other factors may affect future results. Future research could compare exams from courses with large and small (or reverse) gender gaps to further understand how points are allocated on exams with different stakes. We encourage replication studies of the exam retake and stakes phenomena.

VII. CONCLUSIONS

Building from our prior research on this dataset [1], where we found that retake exams were associated with a zero gender gap in introductory courses, we rule out the possibility that this happens because women are more likely to take advantage of the retake opportunity, as men are just as likely to take advantage of the retakes. We also find that the reason for the elimination of the gender gap does not appear to be that women need an extra chance to fail and catch up via practice or extra help (presumably due to any differences in incoming physics preparation) because women appear to do as well as men on their (low-stakes) first exam attempt in the retake classes. These negative results are already important findings because they provide counter evidence to a recurring theme in the literature that gender gaps are due to differences in preparation. The women in the retake classes did not need extra time nor extra tries to perform equally to the men in those classes and so showed no obvious deficiencies.

Instead, we find that women do as well as and perhaps slightly better (though not significantly so) than men on the first, low stakes, exam in a retake class. We also find that women do noticeably better overall on each of the exams (independently true after both first and second tries) in the retake course than they do in the courses without retake exams.

Our data on final exams provide a second, and independent, measurement of an exam gender gap that is inflated by the stakes of the exam. If the stakes of the final exam did not matter for the gender gap, we would find that controlling for in-term exam scores (assessments of the same type, but not the same stakes) would zero the gender gap on the final exams. Instead, we find that controlling for in-term exam scores on the final exams in courses with or without retakes, does not remove the gender gap in final exams, again suggesting that the stakes of the exam control the size of the gender gap in introductory physics exams.

The size of our dataset allows us to measure the extra gender gap on the final exam in these classes to better than one part in ten.

Instructors interested in removing gender performance differences related to exam stakes should reduce the stakes of their assessments by using lower stakes assessment strategies such as implementing retakes on exams.

ACKNOWLEDGMENTS

The authors would like to thank the National Science Foundation for their partial support of this project through the now terminated NSF HSI IUSE 1953760 grant. This work would not have been possible without the support of San José State University's Research, Scholarship, and Creative Activity Assigned Time Program.

DATA AVAILABILITY

The data that support the findings of this article are not publicly available because they contain sensitive personal information. The data are available from the authors upon reasonable request.

APPENDIX A: DETAILED RETAKE EXAM DATA

In the text, we have noted that the class gender gap for in-term exams has a clear dependence on the GPA gender gap in that class. This is shown in Fig. 2, where the class-average in-term exam gender gap is plotted as a function of that class's GPA gender gap for all 254 regular classes. We have also grouped the first attempts of the three retake classes together and plotted them as a single point so that one can see how an average retake class might compare to the nonretake classes. In previous work, we have noted that there seems to be a large exam or grading noise that is not easily explained by differences in student GPAs. Thus, class-dependent gender gaps, which involve differences in noisy grade distributions might be expected to be even noisier. The standard deviation, over classes, of the gender gap is about half of the average gender gap, showing the large variety of results even in classes that have the same instructional materials. Nevertheless, the average retake class sits at the upper edge of the distribution over classes.

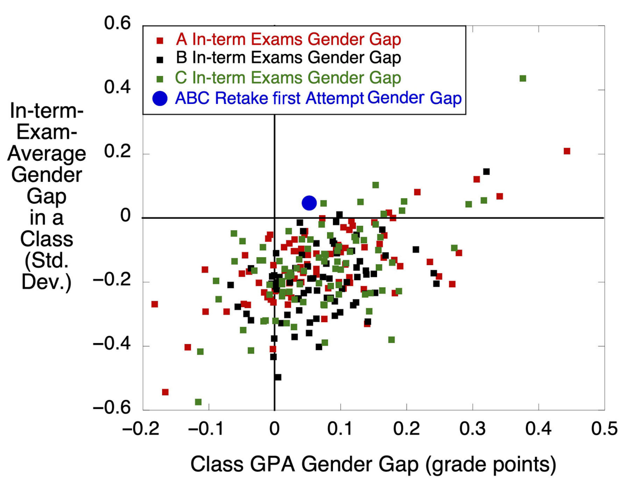

FIG. 2. In-term-exam average z score for female students minus the same for male students plotted as a function of average GPA of female students minus the same for male students for 257 classes in the intro-physics for bioscientists series at one R1 university. The three retake classes for which we have data on the first-attempt exam scores were computed together to show the average effect upon taking an exam for the first time when a retake is possible. No retakes were possible in the other classes so that those exams had higher stakes.

APPENDIX B: INCLUDING A STUDENT ONLY ONCE DOES NOT CHANGE OUR NUMBERS MUCH

As we point out in Sec. III, many of the students in our dataset are represented more than once because they were enrolled in, and received grades in, more than one class included in the data. One worry we had was that including these students more than once might influence our numbers or our conclusions. To check this, we have, for each of the 26 783 unique students in the datafile, randomly chosen only one of their classes to include in an analysis. We made ten of these random choices in order to check the coefficients and errors in Table II. The coefficient resulting from fits with those ten choices were then averaged and are shown in Table IV. We find that the model coefficients we present in the text do not change significantly and that the error estimates change as $\sqrt{1/N}$, where N is the total number of sets of student grades, as one expects if one simply had fewer data points. So the results and conclusions we report in the text are essentially unchanged by having a set of students present in the data more than once because they took more than one class in the series.

Our conclusions from the numerical analysis of final exam gender gaps given in Sec. IV C are also unchanged by

TABLE IV. The average coefficients from ten HLM fits to Eq. (2) (where each fit corresponds to a different random choice of which one class, for each student, is included) are shown along with their standard errors, z statistics, and p values. The comparison with Table II, which includes all classes of all students, shows that the only important effect is to increase the error estimates by the square root of ratio of number of students included in the fit. For this table, the average number of students included (because we know their GPAs) is 24 641 and the retake classes include only 368 students instead of the 57 806 students and 396 students, respectively, included in Table II.

Coefficients	Value	Error	z statistic	p value
b_{GPA}	0.8052	0.0072	112.0	$<10^{-3}$
b_{Female}	−0.2226	0.0072	−31.0	$<10^{-3}$
b_{Retake}	−0.164	0.055	−3.00	0.003
$b_{Female*Retake}$	0.233	0.061	3.85	$<10^{-3}$

including each student only once. When we randomly choose only one appearance for each student in this final exam analysis, we find that the gender gap is that same as shown in Table III, but that the error estimate on the gender gap increases to 0.011 from 0.007 (so z statistic to -10.9 from -16.2).

APPENDIX C: RESULTS FROM THE INTERSECTIONS OF GENDER AND RACE OR ETHNICITY

In our original paper [1], we pointed out that the classes offering retake exams did not have statistically significantly smaller grade gaps between racial or ethnic groups historically marginalized within physics and their peers from groups that have not been marginalized. In the interests of completeness in this paper, we address the intersection of gender with race or ethnicity. We again use Eq. (2) to find the GPA-controlled gender gap difference between the retake and nonretake classes, but now we show the results of 12 separate fits to that model, one for each of the racial or ethnic groups in our database for which both genders are populated in both nonretake and retake classes.

In Fig. 3, we show the results for $b_{Female*Retake}$, as determined from a fit to Eq. (2) for each particular racial or ethnic group described in Table V. One sees that these differences are all positive although most of them are not significantly different than zero. However, it is notable that several are well above zero.

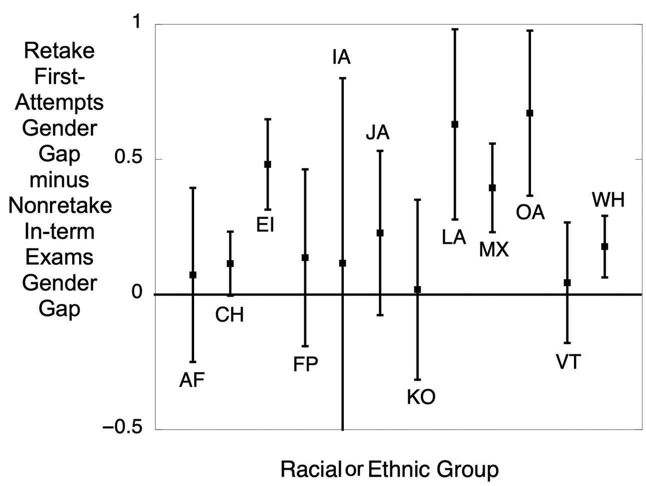

FIG. 3. The, GPA-controlled, gender gap differences between the first-attempt in-term exam average in the retake classes and in-term exams average in the nonretake classes. These are shown for each separate racial or ethnic group for which there are both male students and female students identifying with that racial or ethnic group in both the retake classes and the nonretake classes. We are actually plotting $b_{Female*Retake}$, as determined from a fit to Eq. (2) for each fit to the particular racial or ethnic group described in Table V.

TABLE V. The ethnicities in Fig. 3, as described in the data obtained from the administration of the R1 university of this paper.

Symbol	Ethnicity
AF	African-American/Black
CH	Chinese-American/Chinese
EI	East Indian/Pakistani
FP	Filipino/Filipino-American
IA	Indigenous American/American Indian/ Native American
JA	Japanese-American/Japanese
KO	Korean-American/Korean
LA	Latino/Other Spanish
MX	Mexican-American/Mexican/ Chicano
OA	Other Asian-American/Other Asian
VT	Vietnamese-American/Vietnamese
WH	White/Caucasian

APPENDIX D: GRADE SCALE EFFECT ON GENDER GAP

We have not discussed issues around grade scale and gender in earlier papers, so we will do some of that here. The retake course exams discussed in this paper were graded on a very coarse grade scale. Specifically, an answer that had the physics explained (including setting up equations) fully and correctly received one point (minor arithmetic errors were ignored), an answer that had the physics almost, but not perfectly, complete and correct received 2/3 of a point, and every other answer received zero points. The one-point answers would have received A+ or A or A− under the regular grading in the nonretake classes, and the 2/3-point answers would have received B+ or B or (maybe) B− in the nonretake classes. One might wonder if this kind of coarse grade scale used in the retake classes could explain the lack of a gender gap in the retake classes.

Even though the regular nonretake classes used a variety of grade scales, most of these classes used either a four-point scale or a percent scale so, for these classes, we can identify the fraction of A grades or B grades on specific in-term-exam items received by female students and compare to the fraction received by male students. For this subset of 156 nonretake classes, we find that female students received A grades on $40.02\% \pm 1.3\%$ of their individual exam answers and B grades on $15.24\% \pm 0.07\%$. On these same exams, male students received A grades on $45.05\% \pm 1.7\%$ of their exam answers and B grades on $15.03\% \pm 0.07\%$. So female students, on average, earned about 5% fewer A's and about the same number of B's as male students. If we take these grades from the nonretake classes and rescore all A grades to be one point, all the B grades to be 0.67 points, and the rest of the grades to be zero, then we find the standard deviation of the resulting rescored in-term exam grade distribution is about 0.2. So the average normalized gender gap under this kind of coarse grading would be about (average difference of 0.05 fewer A's) multiplied by (1 point per A) and divided by (standard deviation of 0.2) = 0.25 which, given the crudeness of these grade changes, is reasonably the same as we saw for b_{Female} in Table II. We suggest that the coarse grading had a negligible effect on the gender gap.

APPENDIX E: RETAKE EXAMS GENDER GAP

The group of students taking a retake exam are a self-chosen subset of the full set of students and that subset changes from exam to exam with the exam averages depending on who decides to retake an exam. For these reasons, it is probably unwise to draw any conclusions from analyses of these data. Nevertheless, for completeness, we will calculate the gender gap on the retake exams which we have considered to be medium-stakes exams. Since the group of students retaking an exam changes from week to week, we will just find the gender gap for each week's retake exam and then take a weighted average with the weights determined by the square of the standard error of that exam's gender gap. We, again, z score each retake exam in each class and use HLM to determine the gender gap after controlling for GPA as in Eq. (1). We find that the weighted average of those four retake exam gender gaps is -0.06 ± 0.08. This number is certainly consistent with zero, gap but the 95% confidence interval also overlaps with the other medium-stakes gaps shown in Table II. It is perhaps not surprising that we have no definitive conclusions to draw from these calculations.

APPENDIX F: STUDENTS WHO ARE NONBINARY OR FOR WHOM WE DO NOT HAVE GENDER DATA

Since the university did not collect gender self-identity data until after essentially all of the exams in our dataset had been taken there is very little we can say regarding students who do not identify as one of the two genders that most populate our data. Nevertheless, there were 5 self-identified nonbinary students in our data as well as well as 31 students whose genders were unknown by the university. None of the nonbinary students appeared in a retake class. For completeness, we will extend the final exam analysis we did in Sec. IV C to include both nonbinary students and students whose gender was unknown to the university. We define a categorical variable *Gender* = 0 for students listed as male in our dataset, 1 for those listed as female, 2 for those listed as nonbinary, and 3 for those whose gender is unknown. We extend our final exam analysis to include *Gender* by using HLM to fit

$$FinalExam = b_0 + b_{InTrmExm} InTrmExm + b_{Gender} Gender \tag{F1}$$

where the coefficients of *Gender* each represent the average in-term-exam-controlled final exam difference between that group and the group of male students. All of the coefficients from the HLM fit to Eq. (F1) are given in Table VI. The large p values for the nonbinary and unknown genders suggest that, as expected, there are not enough data to reliably measure differences from male students in final exam scores.

TABLE VI. The coefficients from an HLM fit to final exam z scores [Eq. (F1)] are shown along with their standard errors, z statistics, and p values. Included are 203 classes that had a total of 48 478 students. The coefficients corresponding to the different genders each represent the average difference from the group of male students.

Coefficients	Value	Error	z statistic	p value
$b_{InTrmExm}$	0.9652	0.0051	187.4	$<10^{-3}$
b_{Female}	−0.1157	0.0071	−16.3	$<10^{-3}$
$b_{NonBinary}$	0.27	0.28	0.96	034
$b_{Unknown}$	−0.06	0.12	−0.50	0.62
b_0	0.0744	0.0056	13.3	$<10^{-3}$

[1] D. J. Webb and C. A. Paul, Attributing equity gaps to course structure in introductory physics, Phys. Rev. Phys. Educ. Res. **19**, 020126 (2023).
[2] S. Cotner and C. J. Ballen, Can mixed assessment methods make biology classes more equitable?, PLoS One **12**, e0189610 (2017).
[3] R. R. Valencia, *The Evolution of Deficit Thinking: Educational Thought and Practice* (The Falmer Press, London, 1997).
[4] J. G. Stout, T. A. Ito, N. D. Finkelstein, and S. J. Pollock, How a gender gap in belonging contributes to the gender gap in physics participation, AIP Conf. Proc. **1513**, 402 (2013).
[5] A. Madsen, S. B. McKagan, and E. C. Sayre, Gender gap on concept inventories in physics: What is consistent, what is inconsistent, and what factors influence the gap?, Phys. Rev. ST Phys. Educ. Res. **9**, 020121 (2013).
[6] E. Burkholder and S. Salehi, Exploring the pre-instruction gender gap in physics, PLoS One **17**, e0271184 (2022).
[7] C. Singh and A. Malespina, Test anxiety, self-efficacy, and gender: A quest for equitable assessment practices in

physics, *presented at PER Conf. 2021, Virtual Conference*, 10.1119/perc.2021.pr.Singh.

[8] J. R. Cimpian, T. H. Kim, and Z. T. McDermott, Understanding persistent gender gaps in STEM, Science **368**, 1317 (2020).

[9] 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).

[10] A. Malespina and C. Singh, Gender differences in grades versus grade penalties: Are grade anomalies more detrimental for female physics majors?, Phys. Rev. Phys. Educ. Res. **18**, 020127 (2022).

[11] L. J. Sax, K. J. Lehman, R. S. Barthelemy, and G. Lim, Women in physics: A comparison to science, technology, engineering, and math education over four decades, Phys. Rev. Phys. Educ. Res. **12**, 020108 (2016).

[12] S. Cotner, L. M. Jeno, J. D. Walker, C. Jørgensen, and V. Vandvik, Gender gaps in the performance of Norwegian biology students: The roles of test anxiety and science confidence, Int. J. STEM Educ. **7**, 55 (2020).

[13] S. Salehi, S. Cotner, S. M. Azarin, E. E. Carlson, M. Driessen, V. E. Ferry, W. Harcombe, S. McGaugh, D. Wassenberg, A. Yonas, and C. J. Ballen, Gender performance gaps across different assessment methods and the underlying mechanisms: The case of incoming preparation and test anxiety, Front. Educ. **4**, 107 (2019).

[14] A. Malespina, F. Seifollahi, and C. Singh, Bioscience students in physics courses with higher test anxiety have lower grades on high-stakes assessments and women report more test anxiety than men, Educ. Sci. **14**, 1092 (2024).

[15] T. Espinosa, K. Miller, I. Araujo, and E. Mazur, Reducing the gender gap in students' physics self-efficacy in a team- and project-based introductory physics class, Phys. Rev. Phys. Educ. Res. **15**, 010132 (2019).

[16] R. B. Harris, D. Z. Grunspan, M. A. Pelch, G. Fernandes, G. Ramirez, and S. Freeman, Can test anxiety interventions alleviate a gender gap in an undergraduate STEM course?, CBE Life Sci. Educ. **18**, ar35 (2019).

[17] S. Cwik and C. Singh, Students' sense of belonging in introductory physics course for bioscience majors predicts their grade, Phys. Rev. Phys. Educ. Res. **18**, 010139 (2022).

[18] Z. Hazari, G. Sonnert, P. M. Sadler, and M. Shanahan, Connecting high school physics experiences, outcome expectations, physics identity, and physics career choice: A gender study, J. Res. Sci. Teach. **47**, 978 (2010).

[19] S. Cwik and C. Singh, Damage caused by societal stereotypes: Women have lower physics self-efficacy controlling for grade even in courses in which they outnumber men, Phys. Rev. Phys. Educ. Res. **17**, 020138 (2021).

[20] Y. Li and C. Singh, Effect of gender, self-efficacy, and interest on perception of the learning environment and outcomes in calculus-based introductory physics courses, Phys. Rev. Phys. Educ. Res. **17**, 010143 (2021).

[21] S. Sencar and A. Eryilmaz, Factors mediating the effect of gender on ninth-grade Turkish students' misconceptions concerning electric circuits, J. Res. Sci. Teach. **41**, 603 (2004).

[22] S. R. Wheeler and M. R. Blanchard, Contextual choices in online physics problems: Promising insights into closing the gender gap, Front. Psychol. **10**, 594 (2019).

[23] S. French, A. Dickerson, and R. A. Mulde, A review of the benefits and drawbacks of high-stakes final examinations in higher education, Higher Educ. **88**, 893 (2024).

[24] S.-J. Leslie, A. Cimpian, M. Meyer, and E. Freeland, Expectations of brilliance underlie gender distributions across academic disciplines, Science **347**, 262 (2015).

[25] M. Muradoglu, S. H. Arnold, A. Poddar, A. Stanaland, D. Yilmaz, and A. Cimpian, Why a culture of brilliance is bad for physics, Nat. Rev. Phys. **6**, 75 (2024).

[26] A. Furnham, Self-estimates of intelligence: Culture and gender difference in self and other estimates of both general (g) and multiple intelligences, Pers. Individ. Differ. **31**, 1381 (2001).

[27] A. B. Simmons and A. F. Heckler, Grades, grade component weighting, and demographic disparities in introductory physics, Phys. Rev. Phys. Educ. Res. **16**, 020125 (2020).

[28] J. T. Davis, K. Adams, and A. Morgan, The effect of exam retakes on future exam performance in a large classroom setting, Adv. Physiol. Educ. **48**, 685 (2024).

[29] G. Kortemeyer, W. Bauer, W. Benenson, and E. Kashy, Retaking a test online, Phys. Teach. **44**, 235 (2006).

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

[31] D. J. Webb, C. A. Paul, and M. K. Chessey, Relative impacts of different grade scales on student success in introductory physics, Phys. Rev. Phys. Educ. Res. **16**, 020114 (2020).

[32] C. A. Paul and D. J. Webb, Percent grade scale amplifies racial or ethnic inequities in introductory physics, Phys. Rev. Phys. Educ. Res. **18**, 020103 (2022).

[33] C. Paul, D. J. Webb, M. K. Chessey, and J. Lucas, Pondering zeros: Uncovering hidden inequities within a decade of grades, *presented at PER Conf. 2018, Washington, DC*, 10.1119/perc.2018.pr.Paul.

[34] B. Van Dusen and J. Nissen, Modernizing use of regression models in physics education research: A review of hierarchical linear, Phys. Rev. Phys. Educ. Res. **15**, 020108 (2019).

[35] I. Dar-Nimrod and S. J. Heine, Exposure to scientific theories affects women's math performance, Science **314**, 435 (2006).

[36] C. A. Paul and D. J. Webb, Examining equity and graduation rates at two institutions using a course deficit model and the collaborative learning through active sense-making in physics curriculum, Phys. Rev. Phys. Educ. Res. **21**, 010107 (2025).

[37] J. W. Young and J. L. Fisler, Sex differences on the SAT: An analysis of demographic and educational variables, Res. High. Educ. **41**, 401 (2000).

[38] W. Potter, D. Webb, C. Paul, E. West, M. Bowen, B. Weiss, L. Coleman, and C. De Leone, Sixteen years of collaborative learning through active sense-making in physics (CLASP) at UC Davis, Am. J. Phys. **82**, 153 (2014).