Damage caused by women's lower self-efficacy on physics learning

Z. Yasemin Kalender,[1] Emily Marshman,[2] Christian D. Schunn,[3]
Timothy J. Nokes-Malach,[3] and Chandralekha Singh[1]

[1]*Department of Physics and Astronomy, University of Pittsburgh, Pittsburgh, Pennsylvania 15260, USA*
[2]*Department of Physics, Community College of Allegheny County, Pittsburgh, Pennsylvania 15212, USA*
[3]*Learning Research and Development Center, University of Pittsburgh, Pittsburgh, Pennsylvania 15260, USA*

(Received 11 February 2019; accepted 16 March 2020; published 9 April 2020)

Self-efficacy is an aspect of students' motivation that has been shown to play a critical role in students' engagement, participation, and retention in academic careers in science, technology, engineering, and mathematics (STEM). Since women are underrepresented in STEM domains such as physics, we studied female and male students' self-efficacy and its relation to learning outcomes in physics that can be useful for creating equitable and inclusive learning environments. In a longitudinal study, we surveyed approximately 1400 students in calculus-based physics 2 courses to investigate students' motivational beliefs in physics using a validated survey. We examined female and male students' self-efficacy scores and the extent to which self-efficacy related to learning outcomes (students' grades and conceptual post-test scores), especially the significant gender difference in conceptual post-test scores. To reveal the unique contribution of self-efficacy on outcomes, we controlled for several other variables including Physics 1 grades, SAT math scores, and conceptual pretest scores in physics. We found that the gender differences in conceptual post-test performance were mediated by the model variables. In particular, initial self-efficacy differences showed a direct effect on outcomes even when controlling for students' prior physics knowledge and skill differences, and self-efficacy also had the strongest total gender effect on conceptual learning. Given these findings, future work should focus on better understanding the drivers of these self-efficacy differences including the role that societal stereotypes and biases play in these in order to mitigate these differences.

DOI: 10.1103/PhysRevPhysEducRes.16.010118

I. INTRODUCTION

In the disciplines of science, technology, engineering, and mathematics (STEM), there has been some effort to enhance the participation and advancement of women, yet the historical pattern of overall unequal gender representation remains in many STEM disciplines. Over the past decades, some STEM fields, such as biology and chemistry, have shown great improvement in the number of degrees earned by women [1]. However, other STEM fields, like physics, have seen little progress in increasing representation of women and people of color in the discipline. For instance, the percentage of bachelor's and Ph.D. degrees in physics earned by women in the U.S. is approximately 20% [2]. Even more asymmetric participation occurs for postdoc and academic leadership positions in physics [3].

Education researchers have considered several reasons to explain the gender gap in physics participation [4–13]. These reasons include societal stereotypes and biases pertaining to physics being a discipline for brilliant men [14,15], and related issues such as biased learning tools [4,6], noninclusive teaching methods and physics department climate [5], and motivational factors [7]. Although there has been much interest and research in improving the pedagogy of physics teaching and reforming the content of the physics curriculum, there is relatively less focus on investigating whether these physics learning environments are equitable and how students' motivational factors in calculus-based introductory physics courses (which are foundational courses for many physical science and engineering majors) are related to male and female students' learning of physics. In particular, these motivational factors can lead to differences in performance of male and female students, and can at least partly explain why women do not pursue physics as often as men do.

One of the central motivational factors in educational studies is students' self-efficacy, which refers to individuals' own beliefs about how well they expect to do in a particular subject or task [16]. Prior work in many areas of education has found that self-efficacy predicts students' retention and academic performance even after controlling for knowledge [7,17–19]. Therefore, in understanding gender disparity in physics and to create equitable physics learning environments, self-efficacy is an important factor to examine.

This study examines the role of self-efficacy in explaining the gender gap in college level calculus-based introductory physics courses. At the college level, there are typically fewer than 30% women in calculus-based introductory physics classrooms, compared to algebra-based physics courses in which women are often in the numerical majority [10]. Therefore, it is especially important to understand the role of self-efficacy in explaining gender-based performance gaps in calculus-based physics courses where women are underrepresented. In the following sections, we present an overview of the literature on self-efficacy research in physics learning and its relation to prior knowledge, academic preparation, and performance differences by gender.

II. BACKGROUND

A. Self-efficacy and academic performance

In learning science and educational research, self-efficacy is a commonly used construct that was first proposed by Bandura [20] and it is one of the central factors pertaining to students' beliefs about their capability to perform well in a particular domain [20]. While the impact of societal stereotypes and biases in a discipline on students' self-efficacy can be profound, this motivational factor has been found to shape and be shaped by students' interests, as well as their effort and engagement in class [16]. In particular, self-efficacy can influence students' self-regulation processes, such as goal setting, time management skills, and self-judgement [21]. Students with high self-efficacy become more task centered [22], and they are more likely to exhibit advanced level learning strategies, such as self-monitoring and self-regulation [22]. Likewise, the higher the students' self-efficacy in a particular learning activity, the more perseverance and resilience they are likely to show when faced with adversity [21].

The role of self-efficacy becomes particularly salient when students tackle difficult problems. During problem solving, students with high self-efficacy interpret the struggle as an opportunity for developing their skills while those with low self-efficacy may view the challenge as a large hurdle and further evidence of their lack of competence in the subject [20]. When encountering challenging activities, students with low self-efficacy become less interested, spend less effort and time, and eventually disengage from the class [23]. These behaviors act as a barrier to learning and development.

There is also a strong link between students' self-efficacy and academic performance where low self-efficacy can put students in a negative feedback loop with regard to its impact on performance (which can further lower self-efficacy and negatively impact performance, etc.). In particular, studies in middle and high school have shown that self-efficacy can predict student performance in science courses when controlling for prior knowledge and academic skill differences [24–26]. Relatedly, at the college level, nonphysical science majors' self-efficacy belief was also shown to be a predictor of conceptual understanding and course achievement in physics [27]. In this study, we examine the relationship between self-efficacy and conceptual understanding and course achievement for physical science and engineering majors.

According to Bandura's social cognitive theory, there are several factors that contribute to the development of self-efficacy: mastery experiences (achievement or failure on a previous task), vicarious learning experiences (e.g., observations of how others perform on similar tasks), social persuasion experiences (e.g., cultural norms or biased social messages about who can succeed in a particular domain), and physiological states (e.g., anxiety) [20,28–31]. For instance, having support and encouragement from instructors can positively influence students' self-efficacy and motivate them to engage with difficult learning activities, whereas experiencing stress and doubt due to classroom norms and societal stereotypes might increase disadvantaged students' anxiety and negatively affect their self-efficacy and performance. In this study, we investigate the extent to which students' prior experiences and achievements in math and physics can predict students' self-efficacy and their future physics performance.

B. Self-efficacy, gender, and performance in physics

Many prior studies have shown a prevailing gender gap in students' self-efficacy levels in science and math courses, and in their overall academic achievements. In particular, female students have consistently reported lower self-efficacy than male students in many STEM courses [7,27,32–40]. Cheryan *et al.* investigated causes of gender imbalance in some STEM fields and found that self-efficacy can be a strong predictor of unequal gender participation during class activities and enrollment in STEM fields such as physics [40]. In another study, female students were found to feel less efficacious in physics learning than male students regardless of the type of instruction (i.e., evidence-based active-engagement vs traditional) [34]. Similarly, previous research has identified a large self-efficacy gender gap for equally performing female and male students for all achievement groups (low, medium, high) [12,13]: women who obtained A's in physics on average had self-efficacy levels similar to men who obtained C's.

One of the most well-researched consequences of gender-based beliefs about ability is stereotype threat. In this phenomenon, stigmatized groups such as women in physics have a fear of confirming stereotypical expectations about their gender and they end up performing poorly in physics. In particular, this fear can create further anxiety and can impact the marginalized group's performance (e.g., anxiety can rob students of their cognitive resources while solving problems), which becomes a self-fulfilling prophecy [41]. Although not tested directly in the current study, gender stereotypes provide a well-studied explanation for why physics self-efficacy concerns could lead to differences in learning outcomes.

Previous studies have documented large gender differences in physics performance across various institutions [7–13,42–47]. In college level calculus-based physics courses, women often score lower than men on exams [43] and on standardized conceptual physics tests [44]. Interactive engagement teaching methods have been proposed to address the gender gaps [48]. While some prior work found reduced gender gap in active-engagement courses [47], other studies reported that the performance gap remained [27] and even became larger in calculus-based introductory physics courses despite the use of the interactive teaching methods [46].

Interestingly, the performance gap on standardized conceptual physics assessments has been found to exist on pretests (at the beginning of the course before instruction), which could explain part of the differences in post-test after instruction [44,45]. Therefore, some researchers have suggested that the gender differences in college level physics performance stem from societal stereotypes and biases accumulating over a student's lifetime and the differences between female and male students' high school experiences and preparations [8,49].

Developing robust mathematical skills can help students in college-level physics courses [49]. For example, the number of mathematics courses taken in high school is a strong independent predictor of students' college achievements in introductory science courses [49]. Likewise, research suggests that high school math grades and SAT math scores can predict college physics course success [50–52]. In one study, high school preparation in math was found to be the strongest predictor of students' physics grades in college [8]. Mathematics as a foundation to physics is particularly relevant because there have also been gender differences in math performance [53,54]. Despite female students' high math performance during elementary and middle school, male students score higher on high school math assessments [53,54]. This shift in math achievement during high school might be due to environmental factors such as lack of encouragement for girls in taking more advanced math classes or a belief that math is only for boys due to societal stereotypes and biases [55]. More importantly, gender performance gap in precollege math can further impact women's performance and self-efficacy beliefs in college science courses [56,57].

C. Theoretical framework and research questions

In this study, our primary goal is to explore the mediational mechanism of self-efficacy in explaining gender differences in Physics 2 learning outcomes, while also integrating academic performance (SAT math and Physics 1 course grade) and initial Physics 2 knowledge (standardized conceptual test scores as pretest scores) into the path analysis. We use structural equation modeling (SEM) as an analysis method to unpack the mediational relation between gender and learning outcomes through motivational constructs. SEM is an extension of multiple regression which allows for testing of multiple linear regression models as a single model simultaneously as part of the path analysis; SEM has a number of benefits that are discussed in the methods section. We hypothesize that gender differences in learning outcomes (post-standardized conceptual test scores or course grades) will be mediated by prior knowledge and self-efficacy. Moreover, we also explore the contributions of SAT math and prior knowledge in a standardized conceptual test as additional possible mediators of gender differences in learning outcomes (see Fig. 1). Therefore, our first research question is *To what extent can gender differences in students' physics learning outcomes be explained by differences in physics self-efficacy at the beginning of the course?* Here we contrast the relative roles of self-efficacy, prior knowledge, and SAT math in explaining gender gaps in learning outcomes.

Another important related issue involves the sources of gender differences in physics self-efficacy. Previous work has documented the various ways in which men and women have different exposure to physics within both in and out-of-school learning experiences [58], as well as differential preparation in mathematics. These precollege differences sometimes result in an initial physics knowledge gap and overall mathematics performance gap between men and women when they enter college. These experience differences could also underlie the self-efficacy differences. Therefore, we also posit a second research question: *To what extent are gender differences in physics self-efficacy based on prior physics knowledge differences and measures of pre-college academic performance?* Here we consider Physics 1 grade and SAT math (a common academic skill measure that strongly influences acceptance in selective STEM programs) as a measure of prior knowledge. While those factors are plausible drivers of self-efficacy beliefs, the connections to gender in this population are unclear. In particular, given selective participation of female and male students in physical science and engineering majors, it is not clear in advance whether there are gender differences in Physics 1 grade or SAT math among this population.

Considering all these factors and building on the success of self-efficacy studies in predicting students' achievement and retention, we focus on the impact of self-efficacy across gender on students' college level calculus-based

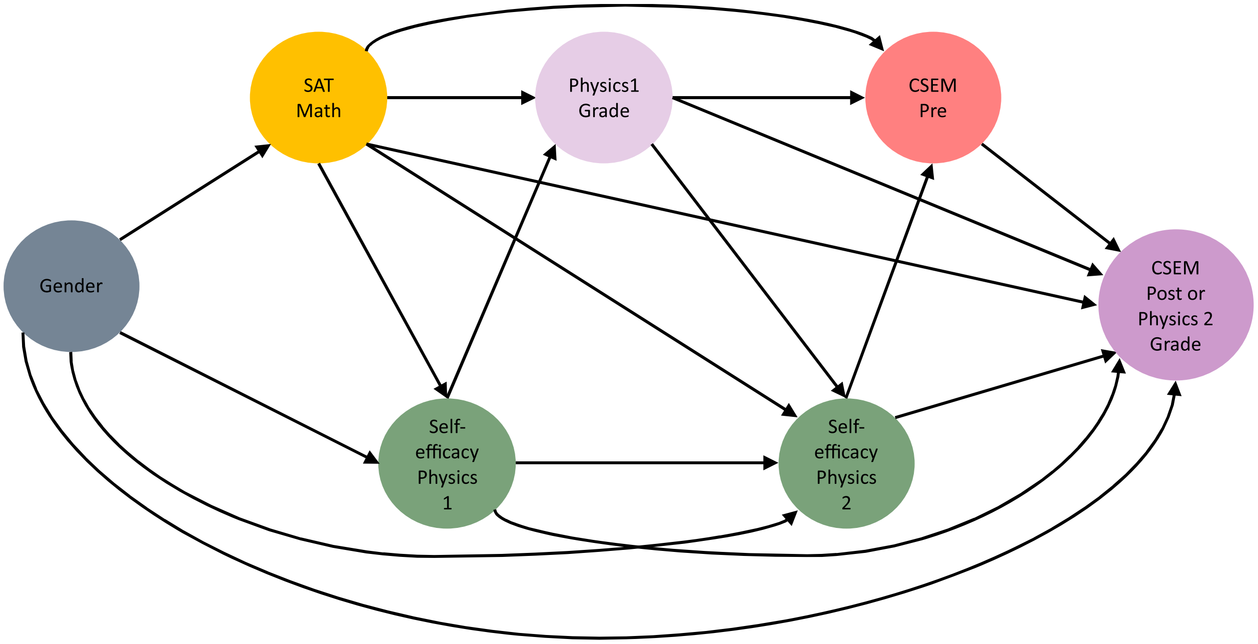

FIG. 1. Conceptual framework connecting gender to learning outcomes (conceptual post-test and Physics 2 course grade) via key college academic experience (Physics 1 grade), attitude (self-efficacy in Physics 1 and 2 at the beginning of the semester in each case), and prior knowledge or skill variables (SAT Math and conceptual pretest CSEM scores). Each arrow corresponds to a linear regression relation between two variables within the path analysis using SEM.

introductory physics performance. Physics is one of the pillar courses taken during the first year of college and it is fundamental to almost all STEM degrees. Positive experiences in first-year physics courses are especially important since students typically decide to stay or exit the major at the end of the first year. Therefore, affirmative first-year experiences in physics courses can play an important role in sustaining female students' interest and self-efficacy in STEM majors [57].

III. METHODOLOGY

Data were collected from introductory calculus-based physics courses over the course of two consecutive years. Our focus is on introductory level Physics 2 courses that encompass topics in introductory electricity and magnetism, which are very challenging topics even for physical science and engineering students partly because they have had relatively little exposure to this specific content in high school. Nationally, many more high schools offer mechanics than electricity and magnetism Advanced Placement (AP) courses, with corresponding large differences in student enrollments (e.g., 2∶1 in calculus-based AP courses and 3∶1 in algebra-based courses in 2019) [59]. Within our sample, the ratio of students who took only mechanics in high school to those who took electricity and magnetism was 3∶1. We examine two different measures of learning outcomes: performance on a research-based standardized conceptual test and course grades.

A. Participants' demographic information and class context

Participants were 1467 students enrolled in calculus-based introductory physics courses, which primarily enroll students who intend to major in engineering or physical sciences. The demographic data (i.e., gender, ethnicity, age) were obtained from the university data warehouse that also kept extensive records about students' pre-college test scores (e.g., SAT) and university grades (e.g., Physics 1 and 2). When motivational and conceptual survey responses were collected, they were sent to an honest broker to be linked with students' demographic information from the university records. Completion of this process gave researchers access to students' survey results merged with their gender and ethnic-racial information as a de-identified dataset.

In terms of demographics, 32% of the students were reported by the university as female; less than 1% of the students had not given gender information and were therefore excluded from this analysis. Although we recognize that gender is a complex sociocultural and multidimensional construct, unfortunately, the data obtained by university records only included binary response options. In future studies, we hope to incorporate gender measurement with multiple options, which can allow us to measure masculinity and femininity in more nuanced ways. Students were predominantly White (78%), with the remaining students coming from a number of other ethnic or racial backgrounds: Asian (12%), African American (4%), Multiracial (3%), Hispanic (2%) and Others (1%). Also, 90% of the students in this course were first-year students with a mean age of 19.

Students in the sample were enrolled in nine sections of Physics 2 courses that were taught by five male White or Asian instructors, having varying levels of teaching experience. To improve generalizability of findings, five of the included sections were taught using traditional lecture-style format and the other four were taught using a flipped class format (i.e., video lectures watched before classed followed by in-class problem solving work). The course topics included electrostatics, magnetostatics, resistance, capacitance, inductance and simple electric circuits, Faraday's law of electromagnetic induction, Ampere-Maxwell's law, Maxwell's equations, electromagnetic waves, and wave optics. There were 24 sections of weekly recitations attached to these lecture sections and they were led by graduate teaching assistants (TA), with women students being a minority in most sections and never a strong majority. All of the TAs were male and slightly less than half were international students. Attendance in recitations was mandatory in that students were given quizzes each week contributing to their final grade.

TABLE I. The physics self-efficacy survey with response options. One item in the survey is reverse coded and indicated as (*R*).

Self-efficacy survey item	Response options
1. I can complete the physics activities I get in a lab class.	a. Rarely b. Half of the time c. Most of the time d. All of the time
2. If I went to a museum, I could figure out what is being shown about physics in [see options to the right]:	a. None of it b. A few areas c. Most areas d. All areas
3. I am often able to help my classmates with physics in the laboratory or in recitation.	a. No!
4. I get a sinking feeling when I think of trying to tackle difficult physics problems. (*R*)	b. no
5. If I wanted to, I could be good at doing physics research.	c. yes
6. If I study, I will do well on a physics test.	d. Yes!

B. Measures

1. *Physics self-efficacy*

We previously developed and validated a self-efficacy survey that was built from prior survey instruments [60–63]. Our instrument was iteratively refined and validated with exploratory factor analysis (EFA), and individual student interviews [10–13]. The individual student interviews used a think-aloud protocol to make sure that students interpreted the questions as intended. Conducting EFAs ensured that items measured self-efficacy coherently and separately from other motivational constructs. Furthermore, we also checked the inter-reliability between the self-efficacy items. In particular, the self-efficacy survey included 6 items and inter-reliability was measured by Cronbach's alpha, where alpha > 0.7 is considered good [64]. Self-efficacy questions assessed students' belief in their ability to understand concepts in physics and their self-perceptions of how they perform certain physics-related activities in and out of the classroom. Table I presents the self-efficacy items and response options for various questions. The main reason for varying response options is to anchor responses in more objective aspect-specific ways and encourage respondents to slow down while responding to read each item (see Table I). Students responded on a scale from 1 (low) to 4 (high), with higher scores indicating higher levels of self-efficacy. Self-efficacy scores were calculated by taking the average responses across the items. For example, a student who answered "all of the time" to the first question, "all areas" to the second question, and "no" to the other four self-efficacy questions would have an average self-efficacy score of $(4 + 4 + 2 + 3$—because this item is reverse coded—$+2 + 2)/(6$ total questions) $=2.83$ which is between positive and neutral (2.5 score) self-efficacy.

We also performed item response theory (IRT) analyses to check the response option distances for survey constructs [65]. These analyses revealed roughly equivalent distance between response items. In particular, the parametric graded response model (GRM) with software STATA was used to test the measurement precision of our response scale [66]. GRM calculates the location parameter for each response and calculates the difference between the locations. The numerical values for the location differences for item responses should be roughly similar in order to support the use of means across ratings [65–67]. The distances for the response options are 2.00 and 2.31, which indicate that it is appropriate to use the averaging of the survey items the way we have done [65–67]. In addition, simple means were highly correlated with IRT factor scores, further justifying the use of means.

2. *Conceptual test*

The Conceptual Survey of Electricity and Magnetism (CSEM) [68] was administered to measure students' conceptual understanding of introductory electricity and magnetism, in contrast to their ability to solve quantitative problems that are typically used in regular course exams (and which can sometimes be solved algorithmically without conceptual understanding of the underlying concepts). The CSEM has been extensively validated as a measure of conceptual understanding of core physics concepts and principles within the course topic areas of electricity and magnetism [68], and it has also been successfully used for comparing different teaching methods on a standardized basis [12,44]. The CSEM test consists of 32 multiple-choice questions. The test was administered at the beginning (pre) and end (post) of the course. We calculated the proportion correct in pre- and post-tests. Typical mean scores on the CSEM for calculus-based physics students (out of 1) was approximately 0.32 on pretest and 0.47 on post-test [12,68]; in other words, the test was very difficult for these students. As is appropriate for scales based upon dichotomous items (e.g., correct or incorrect), we use Armor's θ values to report reliability [69]. The θ values were 0.76 for pretest and 0.84 for post-test, indicating good reliability [69].

3. Course grades

Students' course grades were also used as a measure of their learning outcome. The final course grade was largely determined by students' midterm and final exam scores. Weekly homework, students' class participation, concept quizzes, attendance, and recitation quizzes also contributed to the course grade. The final course grades (both Physics 1 and 2) were obtained from the data obtained from the university records. The conversion between the letter grade and corresponding grade point is given in Table II. While a student's course grade is a measure influenced by attendance, TA and peer support of homework completion, and uneven test quality, this measure is better aligned to full content covered in each course (compared to a standardized test such as CSEM) and also represents an important learning outcome for students (including whether they must repeat the course).

4. Pre-college test scores

The university provided a wide range of scores that are used to determine admission to the university, including high school GPA, standardized assessment scores for mathematical and verbal ability (SAT), and standardized assessment scores for advanced coursework. In our model, we use the Scholastic Assessment Test (SAT) math scores as a predictor variable, which ranged from 400 to 800 and is designed to predict first-year university performance. Prior research suggests that students may overgeneralize the implications of their performance on the SAT, believing that lower math SAT scores imply lower ability for physical sciences [52].

C. Procedures

Motivational and conceptual tests were administered during recitation. Both surveys were administered by the responsible recitation TAs at the beginning of physics courses. The motivational survey was given before students took the conceptual test. The self-efficacy survey was completed by most students in a couple of minutes (embedded in a larger motivational survey taking between 10 and 15 min), and the students worked through the conceptual physics assessments in the remaining class time (approximately 35–40 min).

Instructors were encouraged to give a small amount of course credit to students for completing the surveys. The instructor or teaching assistant responsible for giving the motivational and conceptual physics surveys was given the following script to announce before administering the surveys to the students to encourage students to take the assessments seriously: "We are surveying you on your understanding and beliefs about physics in order to improve the class. Your responses will not be evaluated for grades except to make sure the responses were done seriously, rather than randomly."

D. Analysis

An initial examination compared female and male students' scores in predictors and outcomes for statistical significance using t tests and for effect sizes using Cohen's d [70]. Further, we calculated the correlations between the key constructs for two reasons: highly correlated constructs (>0.90) would signal that they measure nondistinguishable dimensions, whereas low correlations (<0.20) would indicate that the interrelation between the constructs was so low as to not require a direct link in the model (so could be excluded as a variable if not connected to any other variable).

To test the hypothesized path between the variables, we used structural equation modeling (SEM) as a statistical tool by using R (lavaan package) with a maximum likelihood estimation method [71]. SEM is an extension of multiple regression and has multiple advantages compared to other methods. First, by conducting several multiple-regressions simultaneously between variables in one estimation model instead of running them in sequential steps separately, we can calculate the overall goodness of fit and contrast different structural accounts. SEM also enables calculation of interrelated dependence between variables within a single analysis, which has greater statistical power and better controls for indirect correlations through third variables compared to multiple regression models. Third, variables in the model can

TABLE II. Letter grades and corresponding grade points.

	F	*D−*	*D*	*D+*	*C−*	*C*	*C+*	*B−*	*B*	*B+*	*A−*	*A*/*A+*
Grade Point	0	0.75	1.00	1.25	1.75	2.00	2.25	2.75	3.00	3.25	3.75	4
Definitions	Failure					Minimum level to graduate						Superior attainment

be independent variables (input) and dependent variables (output) at the same time, allow for calculation of indirect direct effects through multiple pathways. Finally, SEM has an option to handle missing data by using a full information maximum likelihood "ML" estimation feature, which usually improves both power and generalizability because students missing only some data are not dropped.

SEM involves several commonly used fit parameters to test the goodness of the fit: comparative fit index (CFI), which compares the fit of the proposed model to the null model; Tucker-Lewis index (TLI), which is similar to CFI but takes into account a more complex model—TLI is more strict than CFI; root mean square error of approximation (RMSEA), which refers to residuals and measures how closely the model fit to the data; and standardized root mean square residual (SRMR), which is the standardized difference between the observed correlation and the predicted correlation. There are commonly used thresholds for deciding whether the fit is acceptable or not: CFI and TLI > 0.90; SRMR and RMSEA < 0.08.

Before using mediation as a statistical method, we did moderation analysis to check whether any of the relations between variables show differences across gender or course type (flipped vs traditional). We used the R software package "lavaan" to conduct multigroup SEM. We initially tested for measurement invariance. In other words, we looked at whether the intercepts or residual variances of the observed variables (e.g., self-efficacy, SAT math, etc.) are equal by gender. The analysis involves introducing certain constraints in steps and testing the model differences from the previous step. In each step, we compare the model to both the previous step and the freely estimated model, that is, the model where all parameters are freely estimated for each gender or course type group.

Since we did not find significant moderation by gender or course type (see Appendix), we tested the proposed theoretical model as a mediation analysis, examining the resulting structural paths between constructs. In creating a final acceptable model, we began with the saturated model as shown in Figure 1 (i.e., included all possible regression pathways), and then dropped the connections of variables that were non-significant predictors to obtain a model that produced an acceptable fit to the data and contained only statistically significant regression paths.

Finally, within the path models, the indirect effects of gender to the outcome variables were found by multiplying the coefficients of the particular predictor that connected gender and learning outcome. If the predictor had more than one path between gender and learning outcome, we summed each path's contribution.

IV. RESULTS AND DISCUSSION

A. Correlations

Zero-order pairwise Pearson correlations are given in Table III. Pearson's r values signify the strength of relationship between the variables, uncontrolled for other correlated variables. Investigating the correlations among the predictors (self-efficacy 1 and 2, SAT math, Physics 1 grade, the CSEM pre), we find that there were medium-level correlations around 0.40, showing that the predictors are not so correlated as to be impossible to separate in the regression analyses, but also sufficiently intercorrelated that simple Pearson correlations with outcomes can be artificially higher than the true direct relationships. The strongest correlation was between the students' self-efficacy in Physics 1 and self-efficacy in Physics 2 with $r = 0.59$ (see Table III). The next highest correlation was between students' CSEM pre and Physics 1 Grade ($r = 0.48$) followed by the correlation between Physics 1 grade and Physics 2 self-efficacy ($r = 0.46$). But the $r = 0.46$ correlation represents roughly 20% shared variance, so self-efficacy is not identical to performance measured by this test or necessarily free from biases based upon stereotypes and social interactions. Furthermore, Physics 1 grade was moderately correlated with SAT math test scores, suggesting that prior experience with math is quite important for college level introductory physics courses [43].

The last two rows of Table III present the correlation values between the learning outcomes (the CSEM post and Physics 2 course grade) and the predictors discussed above. The CSEM post-test was most closely correlated with students' Physics 1 grade and CSEM pre-test results.

TABLE III. Pearson intercorrelations are given between all the predictors. Below the thick line shows the correlations between predictors and the learning outcomes (CSEM post and Physics 2 grade). The rightmost column shows the correlation between the two learning outcomes.

		SAT math	Pre Physics 1 self-efficacy	Physics 1 grade	Pre Physics 2 self-efficacy	CSEM pre	CSEM post
Predictors	SAT math	⋯					
	Pre Physics 1 self-efficacy (SE)	0.11	⋯				
	Physics 1 grade	0.43	0.20	⋯			
	Pre Physics 2 self-efficacy (SE)	0.25	0.59	0.46	⋯		
	CSEM pre	0.31	0.26	0.48	0.42	⋯	
Outcomes	CSEM post	0.23	0.32	0.45	0.39	0.40	⋯
	Physics 2 grade	0.28	0.13	0.65	0.30	0.34	0.38

TABLE IV. Means (and standard deviations) of predictor and outcome variables by gender, along with statistical significance (p values after t test) and effect sizes (Cohen's d) for the gender contrast. Theoretical score ranges for each variable are also shown.

	Mean (SD)			
Predictors and outcomes	Female	Male	p value	Cohen's d
SAT Math (400–800)	710 (60)	720 (58)	<0.05	0.17
Pre self-efficacy in physics 1 (1–4)	2.67 (0.46)	2.93 (0.42)	<0.001	0.58
Physics 1 grade (0–4)	2.36 (0.98)	2.48 (1.06)	0.08	0.11
Pre self-efficacy in physics 2 (1–4)	2.58 (0.50)	2.90 (0.45)	<0.001	0.66
CSEM Pre (0–1)	0.36 (0.14)	0.42 (0.14)	<0.001	0.39
CSEM Post (0–1)	0.48 (0.19)	0.55 (0.19)	<0.001	0.40
Physics 2 grade (0–4)	2.39 (0.96)	2.50 (1.04)	0.10	0.10

Both self-efficacy in Physics 1 and 2 courses followed the physics learning results in terms of the correlation value with CSEM Post. For the Physics 2 course grades, students' grades in Physics 1 has the highest correlation value ($r = 0.65$) followed by CSEM post ($r = 0.38$), CSEM pre ($r = 0.34$) and self-efficacy in Physics 2 ($r = 0.30$).

The correlation between the two outcomes variables (CSEM Post and Physics 2 grades) was sizeable but far from identical, supporting the need to separately analyze the relationship of the predictors to the two outcomes. Further, the pattern of simple correlations of Physics 2 course grades and CSEM post with the predictor variables was also different, further suggesting that separate analyses are warranted.

B. Gender differences in predictors and outcomes

Statistically significant gender differences in favor of male students were found on most of the variables (see Table IV). A large gender gap occurred in students' initial self-efficacy reports for both Physics 1 and Physics 2 [70]. While men reported approximately a mean score of 3 in self-efficacy beliefs in both physics courses which corresponded to a positive self-efficacy, women more typically reported a neutral level of self-efficacy (approximate 2.6) in physics at the beginning of the Physics 1 and 2 courses, despite all students being physical science or engineering majors. Further, the gender differences in the standardized performance measures were smaller, with medium differences in CSEM (pre and post), and small differences in SAT math. Thus, while there are preexisting differences based on high school experiences, the largest gender difference appeared to be one of perceived, rather than actual, physics skills or knowledge. The gender gap was smaller in students' course grades than in CSEM performance and it was not statistically significant (see Table IV). Table V shows that female and male students were otherwise very similar in terms of percent underrepresented minorities, intended major, age, and overall GPA at the time of the course, ruling out other demographic differences which might explain performance differences.

C. SEM path model

We used mediation analysis to understand the extent to which gender effects in students' learning outcomes in physics were mediated by differences in students' initial self-efficacy, prior knowledge in physics (as measured by CSEM pre-test and Physics 1 grade), and the pre-college academic measure in math.

1. Using CSEM as a learning outcome

After iterations to remove nonsignificant links, the final mediation model produced good fit parameters: CFI $= 0.99$ (>0.95), TLI $= 0.99$ (>0.95), RMSEA $= 0.02$ (<0.08), and SRMR $= 0.013$ (<0.08). In this model, self-efficacy in Physics 1 and 2, CSEM pretest and Physics 1 grade had direct effect on students' CSEM post scores (see Fig. 2), where there were no direct connections with gender or SAT math. Students' initial Physics 1 grade had the strongest effect ($\beta = 0.40^{***}$) on the conceptual test results. CSEM pre was the second strongest variable that had a direct effect on the CSEM post-test ($\beta = 0.25^{***}$). Finally, self-efficacy scores both in Physics 1 and Physics 2 were the last variables that predict CSEM Post scores.

TABLE V. Key descriptive statistics for female and male students' race or ethnicity, major, GPA, and age.

	Percentage underrepresented minority	Percentage with engineering degree intention	GPA at the end of freshmen year mean (SD)	Average age mean (SD)
Female	13%	68%	3.03 (0.85)	19.1 (1.3)
Male	11%	67%	3.00 (0.75)	19.3 (2.0)

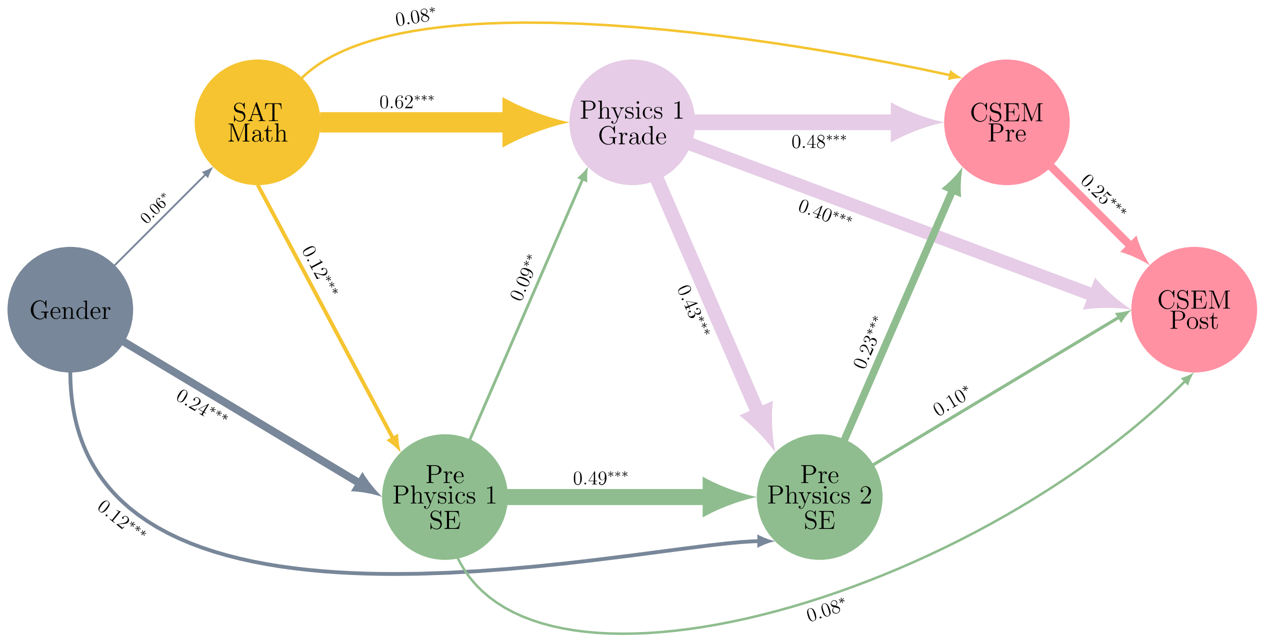

Fig. 2. Results of the structural equation modeling between gender and standardized post-test score (CSEM post) through pre self-efficacy in Physics 1 and 2, SAT math, Physics 1 grade and the CSEM prescore. The line thicknesses correspond to the magnitude of β values. All p values are indicated by *** for $p < 0.001$ and ** for $p < 0.01$. Each arrow with the line connecting two variables in the diagram indicates the direction of regression.

In particular, self-efficacy 1 and 2 remained a significant predictor of learning outcome even after controlling for pre-college academic skills and prior knowledge differences in Physics 1 courses.

More interestingly, we found that students' initial gender differences in Physics 1 courses impact their Physics 1 grade, which later impacts students' Physics 2 self-efficacy with a much larger regression coefficient ($\beta = 0.43^{***}$). The only direct connections to gender involved a relationship with self-efficacy in Physics 1 ($\beta = 0.24^{***}$), self-efficacy 2 ($\beta = 0.12^{***}$), and much small relationship with SAT math ($\beta = 0.06^{*}$). This finding suggests a substantial and powerful impact of students' self-efficacy on learning outcomes even after adjusting for prior knowledge differences: the gender gap in self-efficacy mediated the gendered differences in pre- and post- physics test performance.

2. Using Physics 2 course grade as a learning outcome

For the course grade, a similar model proved to fit the data well, and in fact provided an even stronger fit: $\mathrm{CFI} = 0.99$, $\mathrm{TLI} = 0.99$, $\mathrm{RMSEA} = 0.01$, and $\mathrm{SRMR} = 0.01$. However, there were some structural differences (see Fig. 3). Unlike what we have observed in the first model with CSEM post, Physics 1 grade was only one direct predictor of Physics 2 course grade with a strong connection ($\beta = 0.75^{***}$). The initial gender differences in self-efficacy and SAT Math predicted students' Physics 1 grade which in turn has the strongest direct effect on how students perform on Physics 2 (as measured by grades). The conceptual test (CSEM) mean score in the post-test was 47%. Therefore, Physics 1 grade suppresses the relation between CSEM pre and Physics 2 grade even though there was initially a correlation between two with $r = 0.34$.

D. Total indirect effect of SAT math and Physics 1 self-efficacy

Since gender is not directly connected to either CSEM Post or Physics 2 Grade in the final path models, it is possible to examine the relative contribution of gender to outcomes via SAT math, pre Physics 1 SE, and pre Physics 2 SE since they are measures before students started to interact with college level physics 2 topics. Therefore, we calculated the total mediated effects between gender and learning outcome (Physics 2 grade and CSEM post) via these three variables. Indirect effect of gender to learning outcomes measures the mediated effect by adding all the paths that flow through certain predictor variables after calculating the sum over all the paths which are expressed as the products of β values. We have only counted the paths that had an indirect effect of larger than 0.01 although the pattern is identical when all paths are included. For instance, one of the pre Physics 1 SE mediation paths was gender → pre Physics 1 SE → CSEM post. Therefore, we multiplied all the coefficients along this path ($0.24 \times 0.08 = 0.019$). Another path between gender and CSEM post flowed through variables Physics 1 grade and CSEM pre, so the calculation involved the path: gender → pre Physics 1 SE → Physics 1 grade → CSEM pre → CSEM post. We again multiplied all the standardized

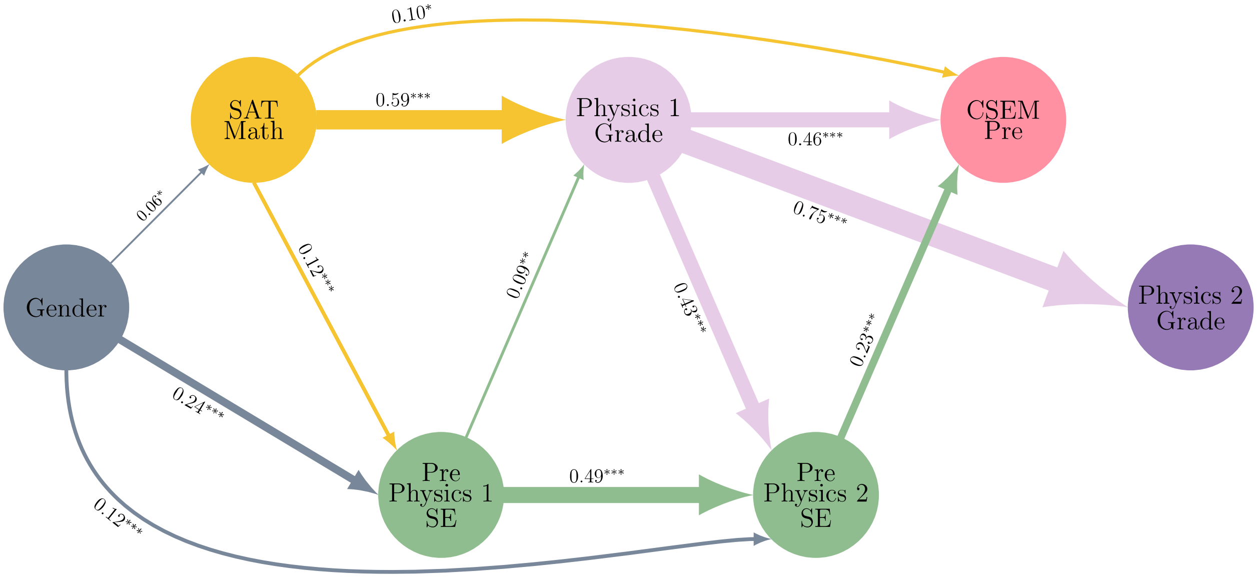

FIG. 3. Results of the structural equation modeling between gender and course grade through pre self-efficacy in Physics 1 and Physics 2, SAT math, Physics 1 grade, and the CSEM pre. The line thicknesses correspond to the magnitude of β values. All p values are indicated by *** for $p < 0.001$ and ** for $p < 0.01$. Each arrow with the line connecting two variables in the diagram indicates the direction of regression.

coefficients for this mediation route as $0.24 \times 0.09 \times 0.48 \times 0.25 = 0.002$. Since this value is very small (smaller than 0.01), this path was not added to the final indirect path calculation. After we calculated all the paths greater than 0.01 that include pre Physics 1 SE in a similar way, we summed them to find the total indirect effect of pre Physics 1 SE. We repeated a similar process for the calculation of SAT math and pre Physics 2 SE as well.

The total mediated effect calculations were conducted separately for both outcome variables (CSEM post and course grade) and shown in Table VI. For the model that has CSEM post as a learning outcome, pre Physics 1 SE had a total indirect effect three times the size of SAT math's total indirect effect. The total indirect effect of pre Physics 2 SE followed the pre Physics 1 SE by half. Therefore, gender was mainly mediated through self-efficacy differences (both in physics 1 and physics 2), which further impacts students' CSEM post. For the second model, where we used Physics 2 grade as a learning outcome, gender's indirect effect was mainly mediated through pre Physics 1 SE followed by SAT math at the half the size. Pre Physics 2 SE had no indirect effect on Physics 2 grade.

TABLE VI. Total indirect gender effects on learning outcomes through SAT math, pre Physics 1 SE, and pre Physics 2 SE. NA indicates nonapplicable.

	Learning outcomes	
Mediator	CSEM post	Physics 2 grade
SAT math	0.015 (1 path)	0.026 (1 path)
Pre Physics 1 self-efficacy	0.043 (3 paths)	0.043 (2 paths)
Pre Physics 2 self-efficacy	0.024 (2 paths)	NA (0 paths)

V. GENERAL DISCUSSION

A. Summary

Our most important finding is that the direct connection between gender and conceptual test results becomes nonsignificant in the SEM model for the conceptual test outcome. Further, the analysis of indirect effects revealed that the gendered patterns in conceptual test performance and course grades were mainly associated with students' self-efficacy, with a smaller role for SAT math. Further, mathematics skills and prior physics preparation appears to be correlated with the large differences in self-efficacy. In particular, prior mathematics and physics learning appears to play a small direct role in shaping later physics learning outcomes but plays an indirect role in shaping physics learning outcomes via undermining or supporting student self-efficacy, which then itself influences learning.

B. Implications

Research suggests that self-efficacy is related to students' learning or performance even after controlling for their prior academic performance differences [12,16,35]. There are several mechanisms that explain the strong impact of self-efficacy on students' motivation, academic achievement, goal orientation, and academic outcome expectations [16–19,72]. Students with high self-efficacy can engage in more challenging tasks without anxiety, which keeps the cognitive load under control, and they are more likely to persist when they face failure in such activities.

In addition to being driven by prior preparation differences, which often result from inequities including societal stereotypes and biases, students' self-efficacy is related to their interactions with peers and classroom experiences [57]. Therefore, in a male-dominated classroom environment such as in calculus-based introductory physics, a woman might experience a lower level of sense of belonging and higher level of anxiety with low self-efficacy [73]. In addition, nonsupportive instructional pedagogies, lack of recognition from instructors, and teaching assistants and classroom interactions with peers can further decrease women's self-efficacy in physics. With that in mind, the instructor's focus on equity and inclusion, and approaches to recognizing students in poorly gender-balanced classrooms become even more vital in supporting women's self-efficacy and promoting learning for all students in the classroom [74]. Since working in an equitable and supportive learning environment will be less likely to trigger stereotype threat, instructors' implementation of explicitly inclusive active-engagement strategies might help women feel more confident and competent in physics. These equitable and inclusive strategies that provide a supportive environment in which women feel recognized and valued might also bolster women's interest in taking more physics-related courses [8,74,75].

Conceptual tests that consist of primarily novel (students are unlikely to have encountered such questions before) and difficult questions is one factor that can activate or elevate stereotype threat and can lower women's self-efficacy and performance. They stand in contrast to traditional exams that comprise more familiar, quantitative questions, that give partial credits for students' solutions rather than only for correctness. As we have shown in this study, the gender gap in conceptual test is mediated by students' self-efficacy.

The gender gap in physics course performance can also increase in active engagement classrooms in which equity and inclusion are not treated as central constructs [46]. Therefore, the reforms towards active-engagement courses in physics not focusing on equity and inclusion will not generally be sufficient on their own to address performance gender gaps. In particular, attention to factors such as equity and inclusivity, the extent to which women feel valued and recognized and details of the support for classroom discussions will be critical in order to benefit all students equally. For example, during classroom activities, instructors must make sure that all students' opinions are valued and respected by all of the group members and all students feel free to communicate without feeling anxious or judged. In group activities during labs or lectures, female students typically have tasks that require a low level of cognitive engagement with the subject matter, such as notetaking or simply reporting the work [73]. Male students might dominate the conversations in these group discussions, which may cause female students' self-efficacy to drop even more. Therefore, in such active-engagement activities, instructors need to assign each student a role and later rotate the student's role and ensure that all students have a sense of belonging and contribute to the task equally.

One primary cause of the gender-biased beliefs in physics is the field-specific intelligence attributions. As Leslie *et al.* found, women recede away from the domains that are thought to require innate ability and brilliance for success in the field [14]. Physics is one of the exemplar fields that illustrates the negative correlation between the number of women and high expectations for brilliance [14]. These biases provoke fixed intelligence mindset attributions regarding how success is achieved via innate ability. Individuals discerning intelligence as a fixed characteristic then perceive struggles as a threat to their ability and failure as an indication of a lack of ability [76]. By contrast, fostering a growth mindset view encourages students to view struggle as a stepping stone that enhances learning, enabling students to become more enthusiastic about spending effort to develop their skills in physics. There are several classroom interventions designed to create better student engagement with growth mindset [77–81]. Some of these interventions have focused particularly on minority groups as they aim to normalize students' struggles in academic life and increase their sense of belonging and self-efficacy [77].

Failure to support women's self-efficacy especially during their first-year college experiences will not only have measurable short-term impact, but is likely to lead to long-term effects, such as gendered patterns of retention in STEM domains. For instance, in the absence of equitable and inclusive learning environments, initial low self-efficacy of women can increase their anxiety in the exams [73] and cause them to perform worse than they actually otherwise would [82]. We also found that conceptual test exams such as CSEM or SAT math had gender gaps whereas we did not find similar gap in students' course grade. While conceptual physics tests are composed of multiple-choice questions and are given in a short period of time (e.g., CSEM tests are given in recitation section and it typically lasts 40–45 min), students' course grades are composed of multiple assessments such as homework, exam grades, participation, etc. Furthermore, physics exams mainly have quantitative problems in both multiple-choice and open-ended formats that give students a chance to obtain partial credit for their solution steps and are made by instructors so that students are more familiar with the types of questions posed. Therefore, we believe that assessments that involve especially difficult and unfamiliar tasks in high stakes' situations can be one factor that elevate stereotype threat especially when these types of assessments are given in a short time. Future work should assess the effect of self-efficacy on exam performance alone as one important component of grades since other measures of grades may mask gender differences on midterm and final exams.

Gender gap in physics self-efficacy favoring male students can have a negative impact on female students' choices of academic career. Some engineering fields are mostly male dominated and contain more physics-focused topics throughout their curricula, while other engineering degrees appear to be more gender balanced and have less focus on physics materials [73]. Because of the first-year college experiences in physics 2 courses, women might switch out from physics-intensive engineering majors, such as mechanical or electrical engineering despite having initial interest in these majors. Equally importantly, due to pervasive societal stereotypes and biases, women's career choices might rely on fixed mindset and the ability-related negative beliefs such as beliefs about women having low ability in physics. Therefore, supporting women in male dominated fields also necessitates promoting and supporting positive recognition and endorsement for their competence from mentors, academic advisors, and course instructors as well as their family members [74]. In particular, academic encouragement, support, and recognition have the potential to enhance their self-efficacy and interest, and help them develop positive identities in the physics-related fields [83–87].

C. FUTURE DIRECTIONS

Future work should involve designing, implementing, and evaluating equitable and inclusive instructional strategies to address the issue of the large gender gap in physics self-efficacy. As discussed earlier, there are some interventions that have been found to improve women's self-efficacy [79–81,88,89] in nonphysics contexts, which may also promote equity and inclusion in the physics classrooms.

ACKNOWLEDGEMENT

This work was supported by Grant No. DUE-1524575 from the National Science Foundation and Grant No. 220020483 from the James S. McDonnell Foundation. Portions of these results were presented at the American Association of Physics Teachers Summer Meeting and Physics Education Research Conference. We would also like to thank Dr. Robert Devaty for his constructive feedback on the manuscript.

APPENDIX: MODERATION ANALYSIS

Using moderation analysis, first, we tested for "strong" or "scalar" measurement invariance by fixing intercepts to equality across gender and course type groups. Both of the models were not statistically significantly different when we compared to freely estimated model. For the gender group, we found chi-square difference $(\Delta\chi^2) = 4.00$, degree of freedom difference $(\Delta\mathrm{df}) = 4$, and nonsignificant $p\,\mathrm{value} = 0.20$ where strong invariance holds. For course type, the strong invariance also holds with, $(\Delta\chi^2) = 4.28$, $(\Delta\mathrm{df}) = 3$ and $p\mathrm{value} = 0.23$. Next step was to test for "strict" measurement invariance where we fixed intercepts and residuals to equality. "Strict invariance" also holds when we compared to the scalar model (gender $\Delta\chi^2 = 7.5$, $\Delta\mathrm{df} = 4$, $p = 0.11$; course type $\Delta\chi^2 = 1.17$, $\Delta\mathrm{df} = 1$, $p = 0.27$) and free model (gender $\Delta\chi^2 = 13.4$, $\Delta\mathrm{df} = 8$, $p = 0.09$; course type $\Delta\chi^2 = 5.4$, $\Delta\mathrm{df} = 4$, $p = 0.24$). Therefore, since strong and strict measurement invariance holds for this model, we continued on to perform other group comparisons.

Next, we ran a multigroup SEM where all regression estimates were fixed to equality for female and male students, and we compared this model to the previous three model (free, scalar and strict). There was no statistically significant difference between two models, so we report the model where regression pathways are equal for men and women. The model fit parameters were good for both gender comparison ($\mathrm{RMSEA} = 0.027$, $\mathrm{SRMR} = 0.04$, $\mathrm{CFI} = 0.989$, $\mathrm{TLI} = 0.985$) and course type comparison ($\mathrm{RMSEA} = 0.028$, $\mathrm{SRMR} = 0.034$, $\mathrm{CFI} = 0.994$, $\mathrm{TLI} = 0.990$). The multigroup SEM results suggest that regression pathways showed very small differences across gender (e.g., from self-efficacy 1 to self-efficacy 2, female students had unstandardized regression coefficient of 0.58 whereas male students had 0.53) or across course type (e.g., from self-efficacy 1 to self-efficacy 2, flipped classroom had unstandardized regression coefficient of 0.59 whereas traditional courses had 0.62). We did not find model differences when we compared this regression model across groups to the freely estimated model (for gender $\Delta\chi^2 = 27.29$, $\Delta\mathrm{df} = 17$, $p = 0.05$; for course type $\Delta\chi^2 = 17.08$, $\Delta\mathrm{df} = 12$, $p = 0.16$); to the scalar model (for gender $\Delta\chi^2 = 21.39$, $\Delta\mathrm{df} = 13$, $p = 0.06$; for course type $\Delta\chi^2 = 12.02$, $\Delta\mathrm{df} = 9$, $p = 0.17$), and to the strict model (for gender $\Delta\chi^2 = 13.8$, $\Delta\mathrm{df} = 9$, $p = 0.12$; for course type $\Delta\chi^2 = 11.02$, $\Delta\mathrm{df} = 8$, $p = 0.15$).

We followed a similar approach when we tested the moderation effect for the second model where we used physics 2 grade as a learning outcome. For gender, we found $\Delta\chi^2 = 9.22$, $\Delta\mathrm{df} = 4$, and nonsignificant $p\,\mathrm{value} = 0.056$. For course type, the strong invariance also holds with $\Delta\chi^2 = 10.42$, $\Delta\mathrm{df} = 5$, and $p\,\mathrm{value} = 0.06$. Next step was to test for "strict" measurement invariance where we fixed intercepts and residuals to equality. "Strict invariance" did not hold when we compared to scalar model and free model for both course type and gender moderation. However, strict invariance rarely holds. Therefore, we continued on to perform regression model comparisons. There was no statistically significant difference between two models, so we report the model where regression pathways are equal for men and women. The model fit parameters were good for both gender comparison ($\mathrm{RMSEA} = 0.015$, $\mathrm{SRMR} = 0.042$, $\mathrm{CFI} = 0.997$, $\mathrm{TLI} = 0.996$) and course type comparison ($\mathrm{RMSEA} = 0.018$, SRMR = 0.038, CFI = 0.997, TLI = 0.996). The multigroup SEM results suggest that regression pathways showed very small differences across gender (e.g., from self-efficacy 1 to self-efficacy 2, female students had unstandardized regression coefficient of 0.54 whereas male students had 0.59) or across course type (e.g., SAT math to Physics 1 grade, flipped classroom had unstandardized regression coefficient of 0.59 whereas traditional courses had 0.62). We did not find model differences when we compared this regression model across groups to freely estimated models (for gender $\Delta\chi^2 = 18.62$, $\Delta\text{df} = 12$, $p = 0.13$; for course type $\Delta\chi^2 = 24.46$, $\Delta\text{df} = 13$, $p = 0.02$); to the scalar model (for gender $\Delta\chi^2 = 9.39$, $\Delta\text{df} = 9$, $p = 0.40$; for course type $\Delta\chi^2 = 14.04$, $\Delta\text{df} = 8$, $p = 0.08$), and to the strict model (for gender $\Delta\chi^2 = 3.58$, $\Delta\text{df} = 5$, $p = 0.61$; for course type $\Delta\chi^2 = 5.76$, $\Delta\text{df} = 7$, $p = 0.56$).

[1] National Science Board Science and Engineering Indicators. National Science Foundation Report, retrieved from website: https://www.nsf.gov/statistics/2018/nsb20181/digest/sections/preface.

[2] American Physical Society, *Bachelor Degrees in Physics and STEM Earned by Women*, retrieved from https://www.aps.org/programs/education/statistics/womenstem.cfm (American Physical Society, 2017).

[3] American Physical Society, *Physics Degrees Earned by Women*, retrieved from https://www.aps.org/programs/education/statistics/womenphysics.cfm (American Physical Society, 2017).

[4] P. V. Engelhardt, in *Getting started in PER*, Reviews in PER Vol. 2, edited by C. Henderson and K. A. Harper (American Association of Physics Teachers, College Park, MD, 2009).

[5] E. Seymour, Tracking the processes of change in US undergraduate education in science, mathematics, engineering, and technology, Sci. Educ. **86**, 79 (2002).

[6] L. McCullough, Presented at the ASQ Advancing the STEM Agenda in Education, the Workplace and Society, University of Wisconsin-Stout, 2011, WWW Document, (http://rube.asq.org/edu/2011/06/continuous-improvement/gender-differences-in-student-responses-to-physics-conceptual-questions-based-on-question-content.pdf).

[7] V. Sawtelle, E. Brewe, and L. Kramer, Exploring the relationship between self-efficacy and retention in introductory physics, J. Res. Sci. Teach. **49**, 1096 (2012).

[8] 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).

[9] R. Ivie and K. Stowe, *Women in Physics 2000 (American Institute of Physics Rep. No. R-430)* (American Institute of Physics, Washington, DC, 2000).

[10] E. Marshman, Z. Y. Kalender, C. Schunn, T. Nokes-Malach, and C. Singh, A longitudinal analysis of students' motivational characteristics in introductory physics courses: Gender differences, Can. J. Phys. **96**, 391 (2018).

[11] T. J. Nokes-Malach, E. Marshman, Z. Y. Kalender, C. Schunn, and C. Singh, Investigation of male and female students' motivational characteristics throughout an introductory physics course sequence, in *Proceedings of the 2017 Physics Education Research Conference, Cincinnati, OH* (2018), pp. 276–279, https://doi.org/10.1119/perc.2017.pr.064.

[12] E. Marshman, Z. Y. Kalender, T. Nokes-Malach, C. Schunn, and C. Singh, Female students with A's have similar self-efficacy as male students with C's in introductory courses: A cause for alarm?, Phys. Rev. Phys. Educ. Res. **14**, 020123 (2018).

[13] Z. Y. Kalender, E. Marshman, T. Nokes-Malach, C. Schunn, and C. Singh, Large gender differences in physics self-efficacy at equal performance levels: A warning sign?, in *Proceedings of the 2018 Physics Education Research Conference, Washington, DC*, (2018), https://doi.org/10.1119/perc.2018.pr.Kalender.

[14] S. J. Leslie, A. Chimpian, M. Meyer, and E. Freeland, Women are underrepresented in disciplines that emphasize brilliance as the key to success, Science **347**, 262 (2015).

[15] A. Maries, N. I. Karim, and C. Singh, Is agreeing with a gender stereotype correlated with the performance of female students in introductory physics?, Phys. Rev. Phys. Educ. Res. **14**, 020119 (2018).

[16] B. Zimmerman, Self-efficacy: An essential motive to learn, Contemp. Educ. Psychol. **25**, 82 (2000).

[17] E. L. Usher and F. Pajares, Sources of self-efficacy in school: Critical review of the literature and future directions, Rev. Educ. Res. **78**, 751 (2008).

[18] M. M. Chemers, L. Hu, and B. F. Garcia, Academic self-efficacy and first year college student performance and adjustment, J. Educ. Psychol. **93**, 55 (2001).

[19] R. W. Lent, B. H. Sheu, D. Singley, J. A. Schmidt, C. L. Schmidt, and C. S. Gloster, Longitudinal relations of self-efficacy to outcome expectations, interests and major choice goals in engineering students, J. Vocat. Behav. **73**, 328 (2008).

[20] A. Bandura, *Self-efficacy*, Encyclopedia of Psychology, 2nd ed., edited by R. J. Corsini (Wiley New York, 1994), Vol. 3, pp. 368–369.

[21] D. Schunk and F. Pajares, *The Development of Academic Self-efficacy*, Development Of Achievement Motivation: A Volume in the Educational Psychology Series edited by A. Wigfield and J. Eccles (Academic Press San Diego, CA, 2002), pp. 15-31.

[22] T. Bouffard-Bouchard, S. Parent, and S. Larivee, Influence of self-efficacy on self-regulation and performance among

junior and senior high-school aged students, Int. J. Behav. Dev. **14**, 153 (1991).

[23] J. M. Bailey, D. Lombardi, J. R. Cordova, and G. M. Sinatra, Meeting students halfway: Increasing self-efficacy and promoting knowledge change in astronomy, Phys. Rev. ST Phys. Educ. Res. **13**, 020140 (2017).

[24] S. Britner and F. Pajares, Sources of science self-efficacy beliefs of middle school students, J. Res. Sci. Teach. **43**, 485 (2006).

[25] S. L. Britner, Motivation in high school science students: A comparison of gender differences in life, physical, and earth science classes, J. Res. Sci. Teach. **45**, 955 (2008).

[26] K. Boden, E. Kuo, T. Nokes-Malach, T. Wallace, and M. Menekse, What is the role of motivation in procedural and conceptual physics learning? An examination of self-efficacy and achievement goals, in *Proceedings of the 2017 Physics Education Research Conference, Cincinnati, OH* edited by L. Ding, A. Traxler, and Y. Cao, pp. 60.

[27] A. Cavallo, W. Potter, and M. Rozman, Gender differences in learning constructs, shifts in learning constructs, and their relationship to course achievement in a structured inquiry, yearlong college physics course for life science majors, School Sci. Math. **104**, 288 (2004).

[28] A. Bandura, Perceived self-efficacy in cognitive development and functioning, Educ. Psychol. **28**, 117 (1993).

[29] A. Bandura, C. Barbaranelli, G. V. Caprara, and C. Pastorelli, Multifaceted impact of self-efficacy beliefs on academic functioning, Child Development **67**, 1206 (1996).

[30] A. Bandura, *Social Learning Theory* (Prentice Hall, Englewood Cliffs, NJ, 1977).

[31] A. Bandura, *Social Foundations of Thought and Action: A Social Cognitive Theory*, Prentice-Hall Series in Social Learning Theory (Prentice Hall, Englewood Cliffs, NJ, 1986).

[32] P. Huang and S. Brainard, Identifying determinants of academic self-confidence among science, math, engineering, and technology students, J. Women Minorities Sci. Eng. **7**, 315 (2001).

[33] S. Brainard and L. Carlin, A six-year longitudinal study of undergraduate women in engineering and science, J. Eng. Educ. **87**, 369 (1998).

[34] J. M. Nissen and J. T. Shemwell, Gender, experience, and self-efficacy in introductory physics, Phys. Rev. ST Phys. Educ. Res. **12**, 1 (2016).

[35] C. Lindstrøm and M. D. Sharma, Self-efficacy of first year university physics students: Do gender and prior formal instruction in physics matter?, Int. J. Innovation Sci. Math. Educ. **19**, 1 (2011).

[36] R. Marra, K. Rodgers, D. Shen, and B. Bogue, Women engineering students and self-efficacy: A multi-year, multi-institution study of women engineering student self-efficacy, J. Eng. Educ. **98**, 27 (2009).

[37] J. Raelin, M. Bailey, J. Hamann, L. Pendleton, R. Reisberg, and D. Whitman, The gendered effect of cooperative education, contextual support, and self-efficacy on undergraduate retention, J. Eng. Educ. **103**, 599 (2014).

[38] A. Zeldin, S. Britner, and F. Pajares, A comparative study of the self-efficacy beliefs of successful men and women in mathematics, science, and technology careers, J. Res. Sci. Teach. **45**, 1036 (2008).

[39] H. Watt, The role of motivation in gendered educational and occupational trajectories related to maths, Educ. Res. Eval. **12**, 305 (2006).

[40] S. Cheryan and V. Plaut, Explaining underrepresentation: A theory of precluded interest, Sex Roles **63**, 475 (2010).

[41] J. W. Osborne, Linking stereotype threat and anxiety, Educ. Psychol. **27**, 135 (2007).

[42] A. T. S. Wee, B. E. Baaquie, and A. C. H. Huan, Gender differences in undergraduate physics examination performance and learning strategies in Singapore, Phys. Educ. **28**, 158 (1993).

[43] P. M. Sadler and R. H. Tai, Success in introductory college physics: The role of high school preparation, Sci. Educ. **85**, 111 (2001).

[44] 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).

[45] 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).

[46] N. I. Karim, A. Maries, and C. Singh, Do evidence-based active-engagement courses reduce the gender gap in introductory physics, Eur. J. Phys. **39**, 025701 (2018).

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

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

[49] R. H. Tai and P. M. Sadler, Gender differences in introductory undergraduate physics performance: University physics versus college physics in the USA, Int. J. Sci. Educ. **23**, 1017 (2001).

[50] G. E. Hart and P. D. Cottle, Academic backgrounds and achievement in college physics, Phys. Teach. **31**, 470 (1993).

[51] D. D. Long, G. W. McLaughlin, and A. M. Bloom, The influence of physics laboratories on student performance in a lecture course, Am. J. Phys. **54**, 122 (1986).

[52] P. Vincent-Ruz, K. Binning, C. Schunn, and J. Grabowski, The effect of math SAT on women's chemistry competency beliefs, Chem. Educ. Res. Pract. **19**, 342 (2018).

[53] D. K. Leonard and J. Jiang, Gender bias and the college prediction of SATs: A cry for despair, Res. High. Educ. **40**, 375 (1999).

[54] H. Wainer and L. S. Steinberg, Sex differences in performance on the math section of the SAT: A bidirectional validity study, Harv. Educ. Rev. **62**, 323 (1992).

[55] C. Benbow and J. S. Stanley, Sex differences in mathematical ability: Fact or artifact, Science **210**, 1262 (1980).

[56] P. Vincent-Ruz and C. Schunn, The increasingly important role of science competency beliefs for science learning in girls, J. Res. Sci. Teach. **54**, 790 (2017).

[57] R. Felder, G. Felder, and E. Dietz, Longitudinal study of engineering student performance and retention. V. Comparisons with traditionally-taught students, J. Engin. Educ. **87**, 469 (1998).

[58] M. G. Jones, A. Howe, and M. Rua, Gender differences in students' experiences, interests, and attitudes toward science and scientists, Sci. Educ. **84**, 180 (2000).

[59] https://secure-media.collegeboard.org/digitalServices/pdf/research/2019/Number-of-Schools-Offering-AP-2019.pdf.

[60] W. K. Adams, K. K. Perkins, N. S. Podolefsky, M. Dubson, N. D. Finkelstein, and C. E. Wieman, New instrument for measuring student beliefs about physics and learning physics: The Colorado Learning Attitudes about Science Survey, Phys. Rev. ST Phys. Educ. Res. **2**, 010101 (2006); B. Zwickl, T. Hirokawa, N. Finkelstein, and H. Lewandowski, Epistemology and expectations survey about experimental physics: Development and initial results, Phys. Rev. ST Phys. Educ. Res. **10**, 010120 (2014).

[61] Activation Lab, Tools: measures and data collection instruments. 2017. Available from http://www.activationlab.org/tools/.

[62] J. Schell and B. Lukoff, Peer instruction self-efficacy instrument, Developed at Harvard University, unpublished Instrument coetta, 2010.

[63] S. Glynn, P. Brickman, N. Armstrong, and G. Taasoobshirazi, Science motivation questionnaire II: Validation with science majors and nonscience majors, J. Res. Sci. Teach. **48**, 1159 (2011).

[64] L. Cronbach, Coefficient alpha and the internal structure of tests, Psychometrika **16**, 297 (1951).

[65] S. E. Embretson and S. P. Reise, *Item Response Theory For Psychologists* (Psychology Press Mahwah, NJ, 2000).

[66] F. Samejima, Estimation of Latent Ability Using a Response Pattern of Graded Scores, Psychometric Monograph, 17(Psychometric Society, Richmond, VA, 1969).

[67] https://www.stata.com/meeting/australia15/abstracts/materials/oceania15_rosier.pdf.

[68] C. Maloney, T. O'Kuma, C. Hieggelke, and A. Van Heuvelen, Surveying students' conceptual knowledge of electricity and magnetism, Am. J. Phys., Suppl. **69**, 12 (2001).

[69] D. J. Armor, Theta reliability and factor scaling, Sociological Methodology **5**, 17 (1973).

[70] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. (Lawrence Erlbaum Associates, Hillsdale, NJ, 1988).

[71] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing (2013), http://www.R-project.org/.

[72] D. H. Schunk, Attributions as Motivators of Self-Regulated Learning, in *Motivation and Self-Regulated Learning: Theory, Research, and Applications*, edited by D. H. Schunk and B. J. Zimmerman (Taylor & Francis, New York, 2008), p. 245.

[73] R. M. Felder, G. N. Felder, M. Mauney, C. E. Hamrin, Jr., and E. J. Dietz, A longitudinal study of student performance and retention. III. Gender differences in student performance and attitude, J. Eng. Educ. **84**, 151 (1995).

[74] Z. Y. Kalender, E. Marshman, C. Schunn, T. N. Nokes-Malach, and C. Singh, Why female science, technology, engineering, and mathematics majors do not identify with physics: They do not think others see them that way, Phys. Rev. Phys. Educ. Res. **15**, 020146 (2019).

[75] Z. Hazari, G. Sonnert, P. 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).

[76] C. Dweck, *Mindset: The New Psychology of Success* (Ballentine, New York, 2006); *Self-theories: Their Role in Motivation, Personality, and Development* (Psychology Press, Philadelphia, 1999).

[77] D. S. Yeager and G. M. Walton, Social-psychological interventions in education: They are magic, Rev. Educ. Res. **81**, 267 (2011).

[78] D. Paunesku, G. M. Walton, C. Romero, E. N. Smith, D. S. Yeager, and C. S. Dweck, Mind-set interventions are a scaleable treatment for academic underachievement, Psychol. Sci. **26**, 784 (2015).

[79] G. M. Walton, C. Logel, J. M. Peach, S. J. Spencer, and M. P. Zanna, Two brief interventions to mitigate a "chilly climate" transform women's experience, relationship, and achievement in engineering, J. Educ. Psychol. **107**, 468 (2015).

[80] D. S. Yeager, G. M. Walton, S. T. Brady, E. N. Akcinar, and D. Paunesku, Teaching a lay theory before college narrows achievement gaps at scale, Proc. Natl. Acad. Sci. U.S.A. **113**, E3341 (2016).

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

[82] P. Pintrich and E. De Groot, Motivational and self-regulated learning components of classroom academic performance, J. Educ. Psychol. **82**, 33 (1990).

[83] G. Potvin and Z. Hazari, The development and measurement of identity across the physical sciences, in *Proceedings of the 2013 Physics Education Research Conference, Portland, OR*, edited by P. V. Engelhardt, A. D. Churukian, and D. L. Jones (AIP, New York, 2013).

[84] A. Godwin, G. Potvin, Z. Hazari, and R. Lock, Identity, critical agency, and engineering: An affective model for predicting engineering as a career choice, J. Eng. Educ. **105**, 312 (2016).

[85] Z. Y. Kalender, E. Marshman, T. Nokes-Malach, C. Schunn, and C. Singh, Beliefs about competence: The story of self-efficacy, gender, and physics, in *Diversity across Disciplines: Research on People, Policy, Process and Paradigm*, edited by A. Murrell and J. Petrie (Information Age Publishing, Charlotte, NC, 2019).

[86] Z. Y. Kalender, E. Marshman, C. Schunn, T. Nokes-Malach, and C. Singh, Gendered patterns in the construction of physics identity from motivational factors, Phys. Rev. Phys. Educ. Res. **15**, 020119 (2019).

[87] M. Good, A. Maries, and C. Singh, Impact of traditional or evidence-based active-engagement instruction on introductory female and male students' attitudes and approaches to physics problem solving, Phys. Rev. Phys. Educ. Res. **15**, 020129 (2019).

[88] P. Haussler and L. Hoffmann, An intervention study to enhance girls' interest, self-concept and achievement in physics class, J. Res. Sci. Teach. **39**, 870 (2002).

[89] J. M. Harackiewicz, E. A. Canning, Y. Tibbetts, S. J. Priniski, and J. S. Hyde, Closing achievement gaps with a utility-value intervention: Disentangling race and social class, J. Personal Soc. Psychol. **111**, 745 (2016).