Sense of belonging is an important predictor of introductory physics students' academic performance

Yangqiuting Li and Chandralekha Singh
Department of Physics and Astronomy, University of Pittsburgh, Pittsburgh, Pennsylvania 15260, USA

(Received 16 January 2023; accepted 23 February 2023; published 3 October 2023)

In this study, we investigated female and male students' sense of belonging and their academic performance measured by the Force Concept Inventory (FCI) and course grades in a calculus-based introductory physics course at a large public university in the United States. We find that compared to male students, female students' sense of belonging was lower on average, and this gender gap persisted from the beginning to the end of the course. Moreover, both women's and men's sense of belonging decreased after the course. We also find that while there was no gender difference in students' Scholastic Aptitude Test (SAT) math scores and female students had a higher average high school grade point average (GPA) than male students, female students had lower average FCI scores and course grades than male students. Using structural equation modeling, we find that students' sense of belonging statistically significantly predicts their FCI scores and course grades after controlling for SAT math scores and high school GPA. Our findings suggest that physics instructors should be intentional about creating an inclusive and equitable learning environment in which students from all demographic groups have a high sense of belonging and can excel.

DOI: 10.1103/PhysRevPhysEducRes.19.020137

I. INTRODUCTION

Prior studies have shown that women are often underrepresented in many science, technology, engineering, and mathematics (STEM) courses and disciplines [1–18]. For example, even though women earn approximately 60% of all bachelor's degrees in the United States, only 20% of the physics undergraduate degrees are earned by women [19]. In addition, some studies have reported gender disparity in students' performance in STEM disciplines [20–24]. Prior research also suggests that individuals' course enrollment and performance in STEM can be influenced by their motivational beliefs such as their sense of belonging in that domain [25–38]. For students from underrepresented groups, a sense of belonging might be undermined due to negative stereotypes about who can excel in certain STEM fields and a lack of encouragement and role models, leading to withdrawal from STEM fields [39–50]. Hence, investigating students' sense of belonging is critical to understanding and addressing diversity, equity, and inclusion issues in STEM disciplines.

Here, by equity in learning, we mean that not only should all students have equitable opportunities and access to learning resources, but they should also have an equitable and inclusive learning environment with appropriate support and mentoring so that they can engage in learning in a meaningful and enjoyable manner and the course outcomes should also be equitable. By equitable course outcomes, we mean that students from all demographic groups (e.g., regardless of their gender identity or race/ethnicity) who have the prerequisites to enroll in the course have comparable course outcomes, which is consistent with Rodrigues *et al.*'s equity of parity model [51]. The course outcomes not only include students' academic performance but also include their motivational beliefs in the domain such as sense of belonging at the end of the course. Sense of belonging not only can influence students' academic performance but it can also play an important role in students' short- and long-term academic and professional goals [52,53]. For example, if students do not believe that they belong in a physics course, they are unlikely to fully engage in learning and are less likely to be enthusiastic about future studies and careers related to physics [52].

A. Students' sense of belonging and academic performance in physics

Sense of belonging in any community is considered one of the basic human needs [54]. In Maslow's original hierarchy of needs with five stages [54], sense of belonging is described as a need for people "to feel like they belong to and are accepted in a social group." Sense of belonging in an academic domain, e.g., physics, is defined as the extent to which students subjectively perceive that they are valued, accepted, and legitimate members of physics [55].

Prior studies have shown that students' academic sense of belonging can influence their self-efficacy, motivation, and intentions to persist in the field [56,57]. In addition, students' sense of belonging can also influence their academic performance [55,58] For example, prior studies have shown that students' sense of belonging in physics class predicts their physics course grades [58,59].

Course grade is a common measure of students' academic outcomes in physics courses. Course grades are usually determined by students' performance on multiple components of the course, such as homework, midterm exams, and final exam. In addition to course grade, physics concept inventories are another common measure of students' physics learning outcomes. Physics concept inventories focus on students' conceptual understanding, and their ability to transfer their conceptual physics knowledge across different contexts and applications [60,61]. Concept inventories are different from typical exams in that their development involves extensive research to ensure their validity and reliability and they allow for standardized comparisons among institutions or over time [61]. Concept inventories are usually given at the beginning and the end of a course to gauge changes in students' understanding. Some studies have shown that students' performance on physics concept inventories is positively correlated to their physics grades [62].

B. Gender difference in students' sense of belonging in physics

Prior studies have shown that women often report a lower sense of belonging in physics class [55,58,59,63]. Some studies show that negative stereotypes about female talent for physics may contribute to the gender difference in physics sense of belonging [58,63]. For example, one study found that female students who endorsed the stereotypes reported a lower sense of belonging in physics, whereas male students were unaffected by endorsing the stereotypes [63]. Moreover, prior studies suggest that the more students believe that innate talent is a prerequisite for success in their field of study, the more likely they are to experience belonging uncertainty [64]. Since the history of physics is often told through the lens of brilliant men from a narrow section of the society, beliefs about "brilliance" can cause female students to doubt their innate ability and can negatively impact their sense of belonging [64]. In addition, the actual gender ratio in the classroom can also influence students' sense of belonging [65]. Women may experience a lower sense of belonging than men in learning environments dominated by men, especially when women are not supported and recognized appropriately by other people in the learning environment [66–69]. In our prior interviews with students, we found that recognition by others, especially from instructors or teaching assistants (TAs) plays an important role in shaping female students' sense of belonging and also their self-efficacy and interest [70–73]. It has been shown that a lack of sense of belonging can lead to withdrawal from physics [58], and prior studies suggest that sense of belonging is more strongly linked to persistence for women than it is for men [53,74]. Therefore, efforts focusing on improving students' sense of belonging, e.g., by developing a more inclusive and equitable learning environment is critical to promote diversity in physics.

C. Gender differences in physics concept inventories and course grades

Prior research suggests that there is often a gender gap favoring men in students' performance on physics concept inventories [75–79]. For example, a study showed that men, on average, outperform women on the mechanics conceptual inventories by 13% on the pretest and by 12% on the post-test [80]. Many studies exploring gender differences in conceptual understanding of introductory classical mechanics have been conducted with the Force Concept Inventory (FCI), which is one of the most commonly used concept inventories for introductory mechanics [81]. For example, McCullough found that the gender gaps on multiple FCI items can be influenced by switching the problem's gender context from stereotypically masculine scenarios to stereotypically feminine contexts [82]. In addition, some other studies show that specific items on the FCI may be biased against women or men. [83,84]. Other factors such as students' academic achievement [85], scientific reasoning ability [86,87], and psychological factors [88,89] have also been analyzed to investigate the gender difference in students' performance on FCI. In addition, some studies suggested that more interactive teaching methods may help reduce the gender gap in students' conceptual understanding [7,90,91]; however, this effect has not been consistently reproduced in other studies [12,20,92]. In particular, a study shows that in a noninclusive learning environment, female students may benefit less from interactive learning because they may not feel safe to express themselves, and thus the gender gap may be even larger than in a traditional lecture-based course [93].

Even though prior studies consistently show gender differences in physics concept inventories, the results of prior studies focusing on gender differences in students' physics course grades or exam scores are less consistent. In particular, while a number of studies indicate that male students outperform female students on physics exams and course grades [59,90], other studies found no significant gender difference in students' performance [77,79,94].

D. Female and male students' precollege academic preparation

Prior studies show that students' performance in college physics courses can be influenced by their precollege academic preparation [95,96]. The Scholastic Aptitude Test (SAT) score and high school grade point average (GPA) are the two most common measures of students' precollege academic preparation. The SAT is a standardized test widely used for college admissions in the United States, which includes two sections (verbal and math). High school GPA is a single cumulative number that represents the student's academic performance in high school.

Prior studies show that women often underperform men on SAT math [59,97–99]. Some studies suggest that the stereotype threat in mathematics tests is one of the factors which contributes to the gender gap [100,101]. Stereotype threat is a situational predicament in which people from a group are at risk of conforming to stereotypes about their social group [14,102]. For example, in math, there exists a societal stereotype that women are less capable than men [15,103]. Studies suggest that stereotype threats can be disruptive enough to impair intellectual performance [14]. For example, in a prior study, gender differences on a math test were found when the female participants believed that the test has shown gender differences before [104].

In contrast to SAT math scores, prior studies show that women who graduated from high school on average have higher high school GPAs than their male counterparts [90,105,106]. Similar results have also been found in smaller sample sizes, such as students in college calculus-based introductory physics courses [90]. Several factors have been proposed in prior studies to explain this phenomenon. For example, high school GPA is the average grade of all courses a student took in high school and is impacted by a number of factors, such as motivation and effort [107,108] as well as attendance and class participation [109,110]. Studies suggest that women typically pay better attention in school and display greater levels of persistence in completing school-related tasks [110,111], which may partially explain the higher high school GPA of female students. In addition, another study suggests that the higher GPA can partly be accounted for by female high school students' higher educational expectations including those for future careers [105]. For example, in comparison with female students, the career plans of male high school students typically included more occupations that do not require high educational achievement and advanced degrees [105].

Prior studies have suggested that students' precollege academic preparation can influence their sense of belonging in college and sense of belonging positively predicts students' physics course grades [58,59]. However, very few studies have been conducted to investigate the effect of sense of belonging on students' physics conceptual understanding and how sense of belonging and conceptual understanding evolve in a physics course. In this study, we focus on students' sense of belonging in a calculus-based introductory physics course. We study how students' sense of belonging predicts their academic outcomes including course grades and performance on the Force Concept Inventory (FCI), which is one of the most well-known and widely used concept inventories in physics for introductory mechanics [81]. We also investigate how female and male students' sense of belonging and FCI scores change from the beginning to the end of the course.

II. RESEARCH QUESTIONS

In this study, we focus on students' sense of belonging and their academic performance in a calculus-based introductory physics course at a large state-related university in the United States. This course is generally mandatory and taken by engineering and physical science majors in the first semester of their first year of undergraduate studies. Our research questions are as follows:

RQ1. Are there gender differences in students' SAT math scores, high school GPA, sense of belonging, FCI scores, and course grades?

RQ2. How do students' sense of belonging and FCI scores change from the beginning (pre) to the end (post) of the course?

RQ3. How does students' sense of belonging predict their FCI scores and course grades after controlling for students' SAT math scores and high school GPA?

RQ4. Does gender moderate the predictive relationship between any two constructs in RQ3? In other words, does the strength of the relationship between any two constructs differ for women and men?

RQ5. If gender does not moderate any predictive relationship between the constructs, how does gender directly or indirectly predict each construct?

We now turn to our theoretical framework for investigating students' sense of belonging and academic performance in the introductory physics course. Based on the theoretical components, we will develop an analytic framework to answer the research questions.

III. THEORETICAL FRAMEWORK

A. Relationship between precollege preparation and performance in physics courses

The SAT score and high school GPA are widely used by colleges to make admission decisions since they have been shown to be predictors of students' success in college [95,96,112]. For example, a prior study suggests that SAT scores and high school GPA can typically account for around 25% of the variance in their college GPA [112]. In another study, SAT math scores explained a significant amount of variance in students' exam performance in a college physics course [95]. Moreover, some recent studies include SAT math scores and high school GPA in a machine learning model to predict students' course grades and help instructors identify at-risk students in physics classes [96]. In addition, SAT math scores and high school GPA have also been found to be positively related to students' performance on physics concept inventories [113,114].

B. Relationship between sense of belonging and academic performance

Prior studies have shown that having a strong sense of belonging in physics courses positively predicts students' course grades [58,59]. In addition to the effect of sense of belonging on students' academic performance, prior research suggests that student performance can also influence students' sense of belonging [115,116]. For example, a study showed that students' sense of belonging after a midterm exam was predicted by their performance on the exam [115]. Precollege academic preparation such as SAT math scores and high school GPA have also been found to predict students' sense of belonging in college physics courses [58,59].

As noted earlier, students' conceptual understanding of physics is another important course outcome and students' performance on physics concept inventories has been shown to predict their physics course grades [62]. In addition, one study suggests that there is a positive correlation between students' sense of belonging and their physics conceptual understanding [117]. However, very few studies have investigated the predictive relationship between students' sense of belonging and their performance on physics concept inventories controlling for gender and high school preparation. Based on the prior studies showing that sense of belonging can predict and be predicted by academic performance [58,115], we hypothesize that sense of belonging can also predict and be predicted by physics concept inventory scores (such as FCI scores).

C. Evolution of students' sense of belonging in physics courses

Students' sense of belonging in physics course can be shaped by multiple factors, for example, students' interaction with their peers and instructors or TAs, quantity and qualities of role models, and endorsement of the stereotypes in physics [55]. Therefore, students' sense of belonging may change from the beginning to the end of the course. Prior studies have shown that female students' sense of belonging decreased after an introductory computing course, while there was no significant change for men [118,119]. Another study showed that both women and men's sense of belonging decreased after a college calculus course [120]. A recent study showed that male students' sense of belonging significantly increased after an algebra-based introductory physics course (mainly taken by bioscience majors), while there was no significant change for women [59]. To our knowledge, no prior studies have investigated how female and male students' sense of belonging evolves in a calculus-based introductory physics course, which is mainly taken by engineering and physical science, and how sense of belonging predicts student FCI performance. Therefore, in this study, we investigate how female and male students' sense of belonging changes from the beginning to the end of a calculus-based introductory physics course and how the sense of belonging predicts students' grades and FCI performance.

IV. THE PRESENT STUDY AND ANALYTICAL FRAMEWORK

Inspired by the research questions and the theoretical components discussed in the last section, we conducted a study focusing on students' sense of belonging and academic performance (including FCI scores and course grades) in a calculus-based introductory mechanics course at a large public university. In this study, we first examined whether there were gender differences in students' high school preparation (including SAT math scores and high school GPA), sense of belonging, FCI scores, and course grades. Then, we investigated how students' sense of belonging and FCI scores changed from the beginning to the end of the course. Next, guided by the theoretical framework, we used structural equation modeling (SEM) [121] to study how students' sense of belonging predicts their FCI scores and course grades after controlling for gender, high school GPA, and SAT math scores. A schematic representation of the SEM model is shown in Fig. 1. In Fig. 1, observed variables are represented by a rectangle and latent variables by an ellipse [121]. In our study, sense of belonging is a latent variable measured by four observed variables (items), which will be discussed in the methodology section. As shown in Fig. 1, high school GPA and SAT math scores are the two high school constructs with a covariance between them, and they predict students' sense of belonging, FCI scores, and course grades as suggested by the theoretical framework. Students' presense of belonging and pre-FCI were measured at the beginning of the course, and their postsense of belonging and post-FCI were measured at the end of the course. There is a regression path from sense of belonging to FCI for both pre and post, and there is also a path from pre-FCI to postsense of belonging. These paths help us test the predictive relationship between students' sense of belonging and FCI scores as we discussed in the theoretical framework section. As suggested by the theoretical framework discussed earlier, course grade is a learning outcome predicted by the other constructs in Fig. 1. Gender is not predicted by any construct, so there is no path pointing to it. The paths from gender to the other constructs capture the gender differences in these constructs after controlling for the effects of the other constructs. In Fig. 1, each construct can be predicted by all the constructs on its left. From left to right, all possible paths were considered in the SEM, but only some of the paths are shown for clarity.

We note that our research design is guided by two epistemological commitments [122]. First, in this study, we focus on how sense of belonging predicts students' FCI scores and course grades. These effects could be mediated by other variables (e.g., students' motivation to learn, engagement in class, interaction with peers and instructors, the level of anxiety, etc.). However, it is still useful to first study the total effect of students' sense of belonging on their academic outcomes. Future studies can investigate

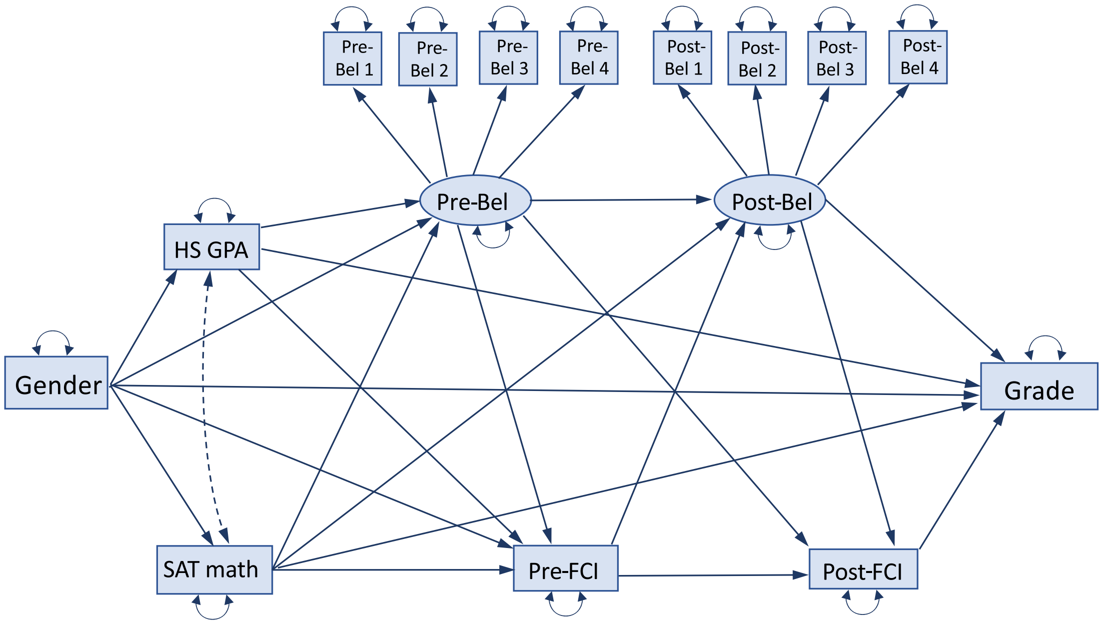

FIG. 1. Schematic representation of the analytic framework. HS GPA represents high school GPA, and Bel represents sense of belonging. The solid lines with a single arrowhead represent regression paths or factor loadings. The curved dashed line with two arrowheads represents a residual covariance. The two-headed curved arrows that exit and reenter the same variable represent the residual variance of the variable.

more complex models with the possible mediators and the role played by each of them in mediating the effect. Second, by using statistical methods such as SEM with a large sample size, our aim is to investigate the relationships between the constructs studied.

V. METHODOLOGY

A. Participants

The data used in this study were collected from students enrolled in a first-semester college calculus-based introductory physics course in Fall 2018 at a large research university in the United States. This course is taken primarily by students majoring in engineering and physical sciences for whom it is mandatory. This course is a traditional lecture-based course (4 h per week) with recitations (1 h per week), in which students typically work on physics problems with the help of a teaching assistant (TA). This course mainly includes mechanics topics such as kinematics, forces, energy and work, rotational motion, gravitation, and oscillations and waves.

The data were collected in the first and last recitation class of the semester. There were 717 students taking the course in the semester studied, and 675 of them participated in our study at the beginning of the course and 638 students at the end of the course. In this study, we focused on 533 students (211 female students and 322 male students) who participated in the study at both the beginning and end of the course (matched students from pre to post). Some possible reasons that some students did not take the presurvey or postsurvey/test include the following: they did not attend the recitations when the survey and test were implemented, or they added or dropped the course after the survey and test were implemented (the add or drop period is the first few weeks of the course). There were no further missing data in our study except for a couple of students forgetting to respond to one survey item related to sense of belonging. In this study, 345 of the students were in engineering school, 174 of the students were in the School of Arts and Sciences, and 14 of the students were in other schools of the university. Of the students who are in the School of Arts and Sciences, only 50 of them had declared their majors when the study was conducted, and only 9 of them were math majors. We note that the average SAT math score is 700 for the math majors and 705 for the other majors, and the difference between them is not statistically significant using independent samples t test ($p = 0.798$). In addition, most students were 18 or 19 years old, and the average age was 18.7 years old with a standard deviation of 1.48. Most students were first-year students, who had just graduated from high school. Students were predominantly White (74%), with the remaining students coming from other ethnic or racial backgrounds: Asian (12%), Hispanic (5%), African American (4%), multiracial (4%), and others (1%). Students' demographic data were provided by the university. We recognize that gender identity is not a binary construct. However, because students' gender information was collected by the university, which offered binary options, we did the analysis with the binary gender data in this study. Fewer than 1% of the

TABLE I. 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 pass						Superior attainment

participants did not provide gender information and therefore were not included in this analysis.

B. Measures

In this study, we investigate how students' sense of belonging predicts their academic performance (measured by FCI scores and course grades) after controlling for students' gender, high school GPA, and SAT math scores. Students' gender identity, high school GPA, SAT math scores, and physics course grades were obtained from the university records. In this study, the gender variable was coded as 1 for male students and 0 for female students. High school grade point average (GPA) and Scholastic Aptitude Test (SAT) math scores are the measures of students' precollege preparation and are widely used for college admissions in the United States. High school GPAs range from 0 to 5 and SAT math scores range from 200 to 800. Students' course grades are largely based on their performance on midterm and final exams, while weekly homework, students' class participation, and recitation practices also contributed to the course grades. The conversion between the letter grade and corresponding grade point is given in Table I. Students' conceptual understanding and their sense of belonging were measured using Force Concept Inventory (FCI) and a validated survey, respectively. Both the conceptual test and survey were administered to students in the first and last recitation class of the semester. Students' performance on the FCI (completeness for pretest and correctness for post-test) was counted as scores for the two corresponding recitations, and the total recitation score is worth a small fraction of the final course grade. We encouraged the instructors to give students a small amount of course credit or extra credit for completing the survey. Students' names and IDs were deidentified by an honest broker who provided each student with a unique new ID. Thus, researchers could analyze students' data without having access to students' identifying information.

1. Conceptual test

The Force Concept Inventory (FCI), which consists of 30 multiple-choice questions, was administered to measure students' conceptual understanding of Newtonian mechanics [81], 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 a conceptual understanding of the underlying concepts). The Force Concept Inventory (FCI) is one of the most commonly used multiple-choice surveys for assessing students' conceptual understanding of Newtonian mechanics [123]. FCI has been validated with extensive interviews and by comparison with the former mechanics' diagnostic test [81,124]. Face and content validity of FCI was established through the support of the numerous physics instructors who have used the test and who have generally agreed that the test measures students' understanding of the force-related concepts [81,124]. In addition, FCI correlates well with other kinematics and force-related assessments such as force and motion conceptual evaluation (FMCE) [80,125]. The reliability of FCI has been well established through extensive use of the test. Hake's large survey data give convincing support for reliability and similar pretest and post-test scores have been found in the United States in many institutions where the style of instruction has been similar [126]. Moreover, FCI shows global test-retest stability [127] and low global context dependence [128].

2. Survey Instruments with belonging items

In this study, we used a validated motivational survey to measure students' sense of belonging at the beginning (pre) and end (post) of the course. There were other constructs in the survey [129–132], but we only focus on the sense of belonging in this study. The survey questions for sense of belonging are listed in Table II. These survey questions were adapted from existing motivational research [132] and were revalidated in our prior work [10,133–136]. The revalidation and refinement of the survey involved the use of one-on-one interviews with students using a think-aloud protocol to ensure that students interpreted the questions as intended [10,66,134,136], exploratory and

TABLE II. Survey questions for sense of belonging, along with CFA factor loadings for students' preresponse and postresponse to the survey. Note that p values are indicated by *** for $p < 0.001$.

No.	Survey items for physics sense of belonging	Lambda Pre	Lambda Post
1	I feel like I belong in this class.	0.681^{***}	0.868^{***}
2	I feel like an outsider in this class.	0.675^{***}	0.793^{***}
3	I feel comfortable in this class.	0.803^{***}	0.817^{***}
4	Sometimes I worry that I do not belong in this physics class.	0.704^{***}	0.831^{***}

confirmatory factor analysis (EFA and CFA) [137], and Cronbach alpha [138,139].

Students' sense of belonging pertains to their feelings of whether they belonged in the physics class [140]. We measured it using four items (shown in Table II) that were scored on a five-point Likert scale: $1 =$ not at all true, $2 =$ a little true, $3 =$ somewhat true, $4 =$ mostly true, and $5 =$ completely true (Cronbach's alpha $= 0.81$ for pre and Cronbach's alpha $= 0.86$ for post). Two items pertaining to sense of belonging ("I feel like an outsider in this class" and "Sometimes I worry that I do not belong in this physics class") were reverse coded to ensure that a higher score in these two items represents a higher sense of belonging. Students' sense of belonging score is the average score of all four items. In Table II, lambda (factor loading) represents the correlation between each item and the sense of belonging construct, and the square of lambda for each item gives the fraction of its variance explained by the construct. As shown in Table II, for both pre and post, all of the CFA item loadings are above 0.5 and most of them are above 0.7, which means that our sense of belonging construct extracts sufficient variance from the items [141].

C. Data analysis

1. Descriptive statistics

In this study, we calculated the mean score for each construct for women and men. We note that sense of belonging was measured using four 5-point Likert scale survey items and each item is a categorical variable. In our previous study [142], we checked the response option distances for the Likert scale items by using item response theory (IRT) to verify the validity of using means across ratings for the sense of belonging construct [143,144]. Even for this study, we performed IRT with the new dataset to verify the validity of using means across ratings. The parametric grades response model (GRM) by using the R software package "mirt" was used to test the measurement precision of our response scale [145,146]. All of the sense of belonging items have the response scale "not at all true, a little true, somewhat true, mostly true, and completely true." GRM calculates the location parameter for each response and calculates the difference between the locations. The results show that the difference between the location parameters were 0.92, 0.94, and 1.15. These results show that the numerical values for the location differences for the item responses are comparable, which suggests that calculating the traditional mean score of the sense of belonging items is reasonable [143,146]. Furthermore, we estimated the IRT-based scores with the expected *a posteriori* computation method for the sense of belonging construct. The results show that the correlation coefficient between the mean score and the IRT-based score is 0.98 for both presense of belonging and postsense of belonging, which also indicates that the use of mean scores is reasonable [143].

Before investigating the gender differences in the constructs studied and the changes in these constructs from the beginning to the end of the course, we first examined the distributions of the collected data, which is important for choosing the appropriate analysis method as suggested by Knaub *et al.* and Springuel *et al.* [147,148]. The distributions of academic data (including high school GPA, SAT math scores, physics course grades, and pre- and post-FCI score) are presented in Fig. 2, and the distributions of female and male students' responses to the Likert scale sense of belonging items are presented in Tables III and IV. The results of the Shapiro–Wilk tests suggest that students' high school GPA, SAT math score, pre- and post-FCI score are not normally distributed. Therefore, we used the Wilcoxon rank-sum test to estimate the gender differences in the constructs studied. The Wilcoxon rank-sum test is commonly used to compare two independent samples when the normality assumption is not satisfied or the data are ordinal [149]. We used the Wilcoxon signed-rank test to estimate the changes in students' sense of belonging and their FCI scores from the beginning to the end of the course. The Wilcoxon signed-rank test is commonly used to compare two matched samples when the normality assumption is not satisfied or the data are ordinal [149].

2. Structural equation modeling

In this study, we used the R [150] software package "lavaan" to conduct structural equation modeling (SEM) [151] to study how students' sense of belonging predicts their FCI scores and grades after controlling for students' gender, high school GPA, and SAT math scores. SEM is a multivariate statistical analysis technique that is used to model the relations between observed variables (items) and latent variables (factors), or between multiple latent variables. This technique is the combination of confirmatory factor analysis (which tests test how well the observed variables represent the latent variables) and path analysis (which estimates the regression relationships between latent variables). Compared with a multiple regression model, a major advantage of SEM is that we can estimate all of the regression links for multiple outcomes and factor loadings for items simultaneously, which improves the statistical power [151]. Another advantage of SEM is that it shows not only the direct regression relation between two constructs but also all the indirect relations mediated through other constructs [151].

The assumptions associated with SEM include correct model specification, sufficiently large sample size, and no systematic missing data [152–154]. The proposed SEM model shown in Fig. 1 is based on the theoretical framework discussed earlier, and each path is suggested or directly supported by prior studies. Therefore, the model specification is theoretically founded. According to Kline, a typical sample size in studies where SEM is used is about 200 [152], so the sample size of our study ($N = 533$) is

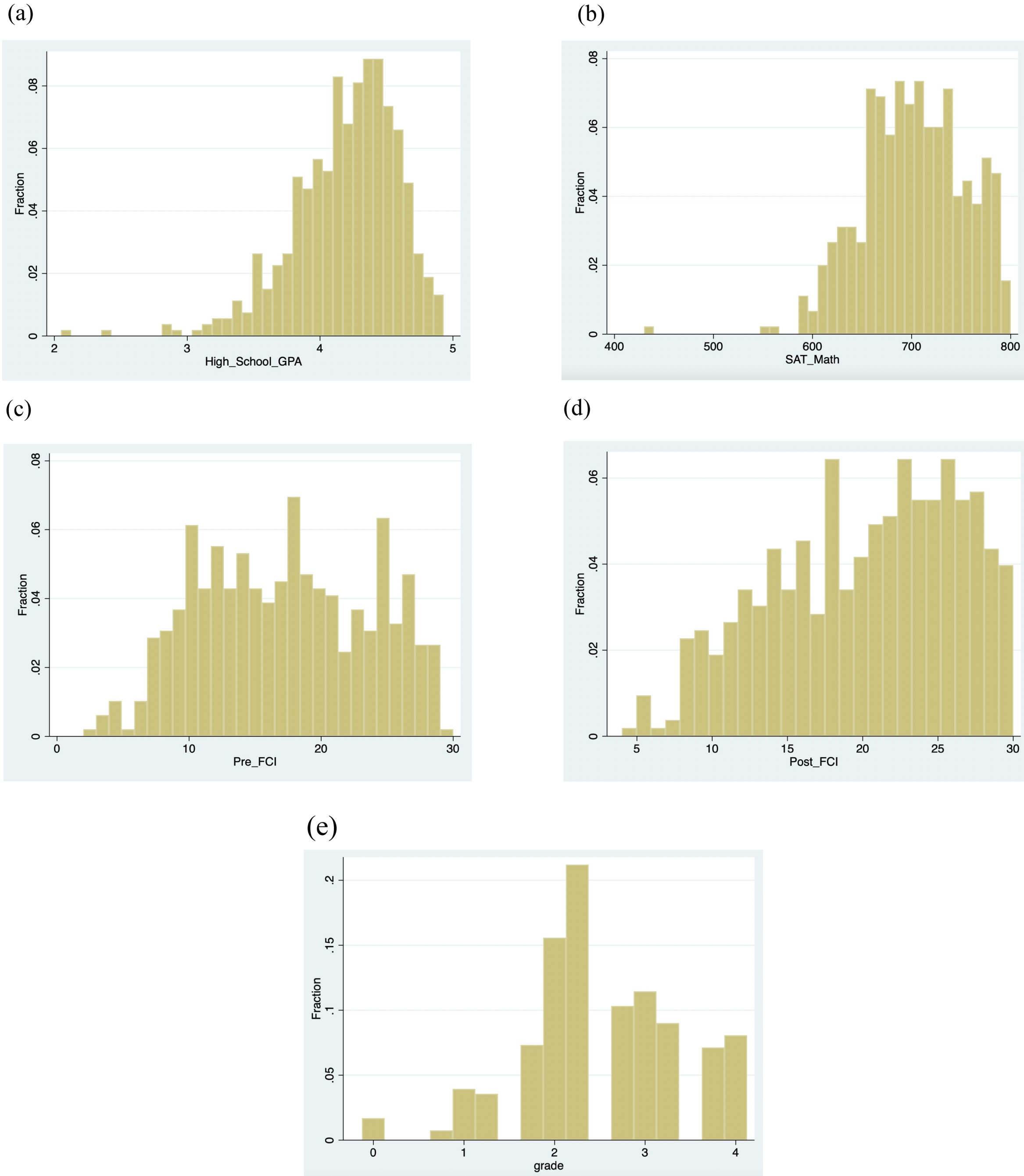

FIG. 2. Graphs of the distributions of (a) high school GPA, (b) SAT math score, (c) pre-FCI scores, (d) post-FCI scores, and (e) physics course grades.

sufficiently large for SEM. Additionally, since we focus on students who participated in both presurvey and postsurvey (matched students from pre to post), students who only participated in presurvey or postsurvey were not included in our analysis. In the results section, we will discuss that the SEM results of using all students' data (unmatched from pre to post) are very similar to the results of using matched data, which suggests that the unmatched students' data do not have systematic missing data. In addition, a well-fitted measurement model (which is also called confirmatory factor analysis) is also very important for performing full SEM [155]. As we will discuss in the next paragraph, our data fit the measurement model very well. Moreover, Table II shows that almost all factor loadings are higher than 0.7, which is considered satisfactory [155]. This means that the constructs extract sufficient variance from the observed variables, which allows us to perform full SEM [156]. In this study, we used diagonally weighted

TABLE III. Percentages of female students who selected each choice from a five-point Likert scale for each survey item of sense of belonging (Bel).

	Pre					Post				
Survey items	1	2	3	4	5	1	2	3	4	5
Bel 1	1%	10%	26%	38%	24%	8%	15%	30%	31%	16%
Bel 2	1%	5%	17%	37%	41%	4%	8%	20%	37%	31%
Bel 3	4%	17%	33%	36%	10%	10%	22%	30%	32%	7%
Bel 4	4%	11%	21%	30%	33%	9%	15%	22%	26%	28%

TABLE IV. Percentages of male students who selected each choice from a five-point Likert scale for each survey item of sense of belonging (Bel).

	Pre					Post				
Survey items	1	2	3	4	5	1	2	3	4	5
Bel 1	1%	4%	27%	37%	31%	4%	8%	30%	34%	23%
Bel 2	1%	3%	8%	31%	56%	3%	8%	13%	33%	43%
Bel 3	2%	10%	32%	40%	16%	5%	17%	33%	29%	16%
Bel 4	3%	3%	13%	36%	45%	5%	10%	18%	31%	36%

least squares (DWLS) to estimate parameters. DWLS estimation is commonly used to analyze ordinal variables and has also been shown to produce unbiased parameter estimates with great statistical power for non-normal data [157,158].

As noted earlier, the SEM includes two parts: confirmatory factor analysis (CFA) and path analysis. First, we performed the CFA for each construct. The model fit is good if the fit parameters are above certain thresholds. In CFA, the comparative fit index (CFI) > 0.9, Tucker-Lewis index (TLI) > 0.9, root mean square error of approximation (RMSEA) < 0.08, and standardized root mean square residual (SRMR) < 0.08 are considered acceptable and RMSEA < 0.06 and SRMR < 0.06 are considered a good fit [159]. In our study, CFI $= 0.976$, TLI $= 0.962$, RMSEA $= 0.056$, and SRMR $= 0.027$, which represents a good fit [159].

Before performing the path analysis, we calculated the pairwise correlations between each pair of constructs (see Table V) [138]. The correlation coefficients were calculated using R software package "lavaan" with DWLS estimator, which is commonly used to estimate correlations between variables when categorical variables are involved [160,161]. In particular, in Table V, the strongest correlation is between pre-FCI and post-FCI with a correlation coefficient of 0.81, which means that students' FCI scores at the end of the course are highly related to their scores at the beginning of the course. In addition, Table V shows that the correlation between high school GPA and students' sense of belonging is not statistically significant at either the beginning or end of the course. This may imply that students with a higher high school GPA do not necessarily have a higher sense of belonging in the physics course.

3. Moderation analysis and model trimming

Since the SEM model in this study involves gender, we conducted a moderation analysis [152,162] to test whether

TABLE V. Pairwise correlation coefficients of the constructs. Note that p values are indicated by *** for $p < 0.001$, ** for $0.001 \leq p < 0.01$, * for $0.01 \leq p < 0.05$, and ^{ns} for $p > 0.05$ (not statistically significant).

Variable	1	2	3	4	5	6	7
1. SAT math	…	…	…	…	…	…	…
2. High school GPA	0.28^{***}	…	…	…	…	…	…
3. Prebelonging	0.16^{**}	0.00^{ns}	…	…	…	…	…
4. Postbelonging	0.31^{***}	0.09^{ns}	0.59^{***}	…	…	…	…
5. Pre-FCI	0.34^{***}	0.10^{*}	0.37^{***}	0.49^{***}	…	…	…
6. Post-FCI	0.37^{***}	0.15^{***}	0.36^{***}	0.48^{***}	0.81^{***}	…	…
7. Grade	0.49^{***}	0.38^{***}	0.28^{***}	0.50^{***}	0.54^{***}	0.60^{***}	…

gender moderates the relationship between any two constructs in the model (i.e., do the strength of relationships given by the standardized regression coefficients between any two constructs in the model differ for women and men?). We used the R [150] software package "lavaan" to conduct multigroup SEM. We initially tested for measurement invariance. In other words, we analyzed whether the factor loadings, intercepts, and residual variances of the observed variables are equal in the model where we measured the latent constructs so we can confidently perform multigroup analysis. The analysis involved introducing certain constraints in steps and testing the model differences from the previous step. In each step, we compared the model to both the previous step and the freely estimated model, that is, the model in which all parameters are freely estimated for each gender group. First, to test for "weak" or "metric" measurement invariance, we ran the model where only factor loadings were fixed to equality across both gender groups, but intercepts and errors were allowed to differ. The model was not statistically significantly different from the freely estimated model according to a likelihood ratio test, so weak measurement invariance holds (chi-square difference $(\Delta\chi^2) = 8.817$, degree of freedom difference $(\Delta\text{d.o.f.}) = 6$, and nonsignificant $p = 0.184$). Next, we tested for "strong" or "scalar" measurement invariance by fixing both factor loadings and intercepts to equality across gender groups. This model was not statistically significantly different from either the metric invariance model ($\Delta\chi^2 = 6.977$, $\Delta\text{d.o.f.} = 6$, $p = 0.323$) or the freely estimated model ($\Delta\chi^2 = 15.794$, $\Delta\text{d.o.f.} = 12$, $p = 0.201$), so strong measurement invariance holds. Finally, to test for "strict" measurement invariance, we fixed factor loadings, intercepts, and residual variances to equality. This model was not statistically significantly different from either the scalar invariance model ($\Delta\chi^2 = 7.892$, $\Delta\text{d.o.f.} = 8$, $p = 0.444$) or the freely estimated model ($\Delta\chi^2 = 23.685$, $\Delta\text{d.o.f.} = 20$, $p = 0.256$), so strict measurement invariance holds. Therefore, since all levels of measurement invariance hold for this model, we continued to perform multigroup comparisons [152,162].

We ran a multigroup SEM in which all regression estimates were fixed to equality for female and male students in addition to the factor loadings and intercepts, and we compared this model with the freely estimated model. There was no statistically significant difference between the two models, so we reported the model in which regression pathways are equal for men and women. The model fit parameters for this case were acceptable ($\text{RMSEA} = 0.053$, $\text{SRMR} = 0.065$, $\text{CFI} = 0.962$, and $\text{TLI} = 0.960$). The multigroup SEM results suggest that regression pathways among the constructs do not have differences across gender when compared to the freely estimated model ($\Delta\chi^2 = 44.479$, $\Delta\text{d.o.f.} = 34$, $p = 0.108$) or to the strict model ($\Delta\chi^2 = 20.794$, $\Delta\text{d.o.f.} = 14$, $p = 0.107$). Therefore, the results shown above indicate that in our model, strong measurement invariance holds and there is no difference in any regression coefficients by gender, which allowed us to perform the path analysis involving gender using SEM [152,162] (as shown schematically in Fig. 1).

We first analyzed the saturated SEM model that includes all possible links from left to right between different constructs shown in Fig. 1, and then we removed the most insignificant path line (with the highest p value) and reran the model. We used this method to trim one path at a time until all remaining path lines were statistically significant [162]. Next, we used modification indices to improve the model fit. The modification index is the chi-square value, with one degree of freedom, by which model fit would improve if a particular path was added back [162]. A modification index larger than 3.84 indicates that the model fit would be significantly improved, and the p value for the added parameter would be <0.05 [163,164]. We added back the paths with a modification index larger than 3.84 one at a time (from high to low modification index) to improve the model fit. Finally, we checked the statistical significance of each trimmed path by adding them back to make sure that all trimmed paths are not statistically significant and that all statistically significant paths are kept.

VI. RESULTS

A. Descriptive statistics for students' sense of belonging and academic performance

Table VI shows the descriptive statistics of students' sense of belonging and FCI scores, along with the results of Wilcoxon rank-sum tests for gender differences and Wilcoxon signed-rank tests for changes from the beginning to the end of the course. Cohen suggested that typically values of 0.1, 0.3, and 0.5 represent small, medium, and large effect sizes, respectively, e.g., for Wilcoxon rank-sum tests and Wilcoxon signed-rank tests [165]. Hake suggested that values of $g < 0.3$, $0.3 < g < 0.7$, and $g > 0.7$ represent small, medium, and large normalized gains, respectively [126]. As shown in Table VI, female students had statistically significantly lower scores in both sense of belonging and FCI at the beginning of the course, and the gender differences were maintained at the end of the course. In addition, we found that from pre to post, both female and male students' FCI scores increased, and the effect size and normalized gain indicate that female and male students have a similar amount of improvement in FCI scores. However, Table VI shows that both female and male students' sense of belonging decreased from pre to post, and the effect size indicates a similar amount of drop for women and men.

Table VII shows students' high school GPA, SAT math scores, and grades in the physics course. As shown in Table VII, there was no statistically significant gender difference in students' SAT math scores, and female

TABLE VI. Descriptive statistics of presense and postsense of belonging (Bel) and FCI scores for female and male students. A minus sign indicates that students' average score decreased from pre to post.

	Pre-Bel	Post-Bel	Statistics	
Gender	Mean	Mean	Effect size	p value
Male	4.01	3.71	−0.23	<0.001
Female	3.74	3.42	−0.22	<0.001
p value	<0.001	0.001		
Effect size	0.16	0.14		

	Pre-FCI	Post-FCI	Statistics		
Gender	Mean	Mean	Normalized gain (g)	Effect size	p value
Male	63%	72%	0.23	0.45	<0.001
Female	50%	61%	0.23	0.48	<0.001
p value	<0.001	<0.001			
Effect size	0.29	0.26			

TABLE VII. Descriptive statistics of female and male students' high school GPA, SAT math scores, and course grades. A minus sign indicates that female students have a higher average score than male students. "ns" indicates that the effect size is not statistically significant.

	Mean			
Grades (score range)	Male	Female	p value	Effect size
High school GPA (0–5)	4.12	4.31	<0.001	−0.23
SAT math (200–800)	706	703	0.612	ns
Grade (0–4)	2.57	2.44	0.033	0.09

students had a higher average high school GPA than male students. In addition, women have slightly lower average course grades than men.

B. Structural equation modeling

In this section, we discuss the use of structural equation modeling (SEM) to investigate how students' sense of belonging predicts their course grades and FCI scores after controlling for students' gender, high school GPA, and SAT math scores. As noted earlier, our moderation analysis shows that gender does not moderate the relationship between any two constructs studied, so we can include gender as a variable in the SEM model as shown in Fig. 1. Figure 3 shows the results of the SEM model. The chi-square statistic for this model is $\chi^2(68) = 119.864$, $p < 0.001$, which suggests that there is a statistically significant difference between the covariances predicted by the model and the population covariance matrix [152]. However, prior studies suggest that χ^2 is sensitive to sample size; as the sample size increases, the χ^2 test statistic has a tendency to indicate a significant probability [166]. As a result, the significance of χ^2 becomes less reliable when sample sizes are outside a range (generally above 200) [166]. Therefore, the level of SEM model fit is often represented by the comparative fit index (CFI), Tucker-Lewis index (TLI), root mean square error of approximation (RMSEA), and standardized root mean square residuals (SRMR), and $\mathrm{CFI} > 0.9$, $\mathrm{TLI} > 0.9$, $\mathrm{RMSEA} < 0.08$, and $\mathrm{SRMR} < 0.08$ are considered acceptable [159]. The model shown in Fig. 3 fits the data well with $\mathrm{CFI} = 0.980$, $\mathrm{TLI} = 0.977$, $\mathrm{RMSEA} = 0.044$, and $\mathrm{SRMR} = 0.036$ [159].

As shown in Fig. 3, students' sense of belonging at the end of the course directly predicts their grades even after controlling for their pre-FCI, SAT math, and high school GPA. In particular, the direct effect of postsense of belonging on grades is $\beta = 0.20$, which is comparable to the effects of SAT math and high school GPA on students' grades. In addition, students' presense of belonging directly predicts their pre-FCI with $\beta = 0.25$. We note that gender predicts high school GPA with a negative regression coefficient ($\beta = -0.23$), while gender predicts presense of belonging and pre-FCI with positive regression coefficients ($\beta = 0.20$ for presense of belonging and $\beta = 0.27$ for pre-FCI). Since the gender variable was coded as 1 for male students and 0 for female students, paths from gender with $\beta > 0$ and $\beta < 0$ indicate a higher mean for men and women, respectively, in the predicted variable. Therefore, female students on average had a higher high school GPA but a lower sense of belonging and FCI scores than male students. These results are consistent with the descriptive statistics shown in Tables VI and VII. We note that even though there are statistically significant gender differences favoring men in postsense of belonging, post-FCI, and grade, gender does not directly predict any of them. This result means that the gender differences in these constructs are mediated through the other constructs in the model, such as presense of belonging and pre-FCI, which are directly predicted by gender. We note that the results shown in Fig. 3 are based on data of the students matched from pre

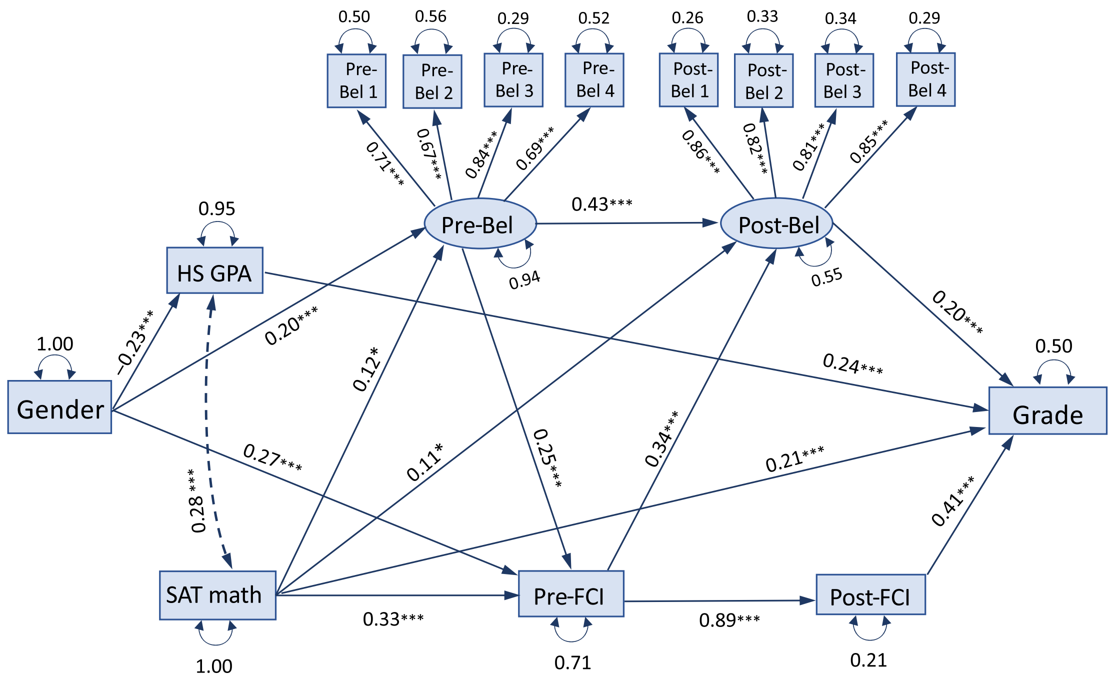

FIG. 3. Results of the SEM model. HS GPA represents high school GPA, and Bel represents sense of belonging. The solid lines with a single arrowhead represent regression paths or factor loadings. The curved dashed line with two arrowheads represents a residual covariance. The two-headed curved arrows that exit and reenter the same variable represent the residual variance of the variable. The numbers on the paths between constructs are regression coefficients (β values), which represent the strength of the regression relations, and p values for the regression coefficients are indicated by *** for $p < 0.001$, ** for $0.001 \leq p < 0.01$, * for $0.01 \leq p < 0.05$.

to post. In the Appendix, we present the results of the SEM model using all students' data (unmatched from pre to post), which are very similar to the results presented in Fig. 3.

To further understand how the model explains the variance in each construct, we calculated the coefficients of determination R^2 (fraction of variance explained) for each construct. We found that R^2 values of all outcome constructs are reasonably high. In particular, $R^2 = 0.79$ for post-FCI, $R^2 = 0.50$ for course grades, and $R^2 = 0.45$ for postsense of belonging. This means that our models explain much variance in the outcome constructs. We note that presense of belonging ($R^2 = 0.06$) and pre-FCI ($R^2 = 0.29$) have relatively small R^2 values since they are predicted by fewer constructs.

VII. SUMMARY AND DISCUSSION

In this study, we focused on students' sense of belonging and academic performance in a college calculus-based introductory physics course. In particular, we first studied the gender differences in students' high school preparation, sense of belonging, and academic performance in the course. Then, we investigated how students' sense of belonging and FCI scores change from the beginning to the end of the course. Finally, we studied how students' sense of belonging predicts their FCI scores and course grades after controlling for their gender, high school GPA, and SAT math scores.

In response to RQ1, our results show that women have a higher average high school GPA than men, which is consistent with prior studies. However, we found that there was no statistically significant gender difference in students' SAT math scores, which is different from the prior studies showing that women underperformed compared to men on SAT math in other contexts [59,97–99]. We note that most prior studies focusing on SAT math sampled high school students taking the SAT. Prior studies show that students consider college admission as a motivation to take the SAT, and there are more women taking the SAT and enrolling in college than men in the United States [167,168], which means that the sample of men taking the SAT may be more selective than women [99]. Our study focuses on students in a calculus-based introductory physics course, in which men far outnumber women. This difference in sample composition between our study and the prior studies may partially explain the difference in the findings regarding gender differences in SAT math scores. We note that in our study, even though women have comparable (SAT math scores) and even better precollege academic preparation than men (high school GPA), women have statistically significantly lower sense of belonging and FCI scores in the physics course (RQ1). This result is consistent with prior studies showing the gender disparity in sense of belonging and FCI scores [17,55].

In response to RQ2, we found both women and men's FCI scores increased from the beginning to the end of the physics course, which indicates that this course indeed helped students improve their conceptual understanding of mechanics. However, we found that both women and men's sense of belonging decreased after the course. The decrease in students' sense of belonging has also been found in prior studies conducted in other courses such as introductory computing course and calculus course [119,120]. These findings may partly be related to the "weed-out" culture in physics and some other STEM disciplines [53,55]. Many mandatory introductory courses in STEM are large in enrollment and challenging, so in the absence of efforts to create equitable and inclusive learning environments, they appear to be gatekeepers and weed out students who perform poorly. This weed-out culture might impact students' sense of belonging, especially for students from underrepresented groups such as women, who already have a lower sense of belonging than students from the majority groups on average. In addition, most students in our study were in the first semester of their first year of undergraduate studies. Therefore, they might experience some uncertainty and anxiety during the transition from high school to college, which may also influence their sense of belonging. Moreover, the inclusiveness of the learning environment may also influence students' sense of belonging [169]. For example, in our prior qualitative studies conducted in the calculus-based physics courses, we found that students' sense of belonging was negatively impacted when their questions were not valued by the instructors or when the class was dominated by a small group of students [66,170].

In this study, we note that both women and men's sense of belonging and FCI scores changed from the beginning to the end of the course, and the effect sizes or normalized gains for these changes are almost the same for women and men. As a result, the gender differences in these constructs are maintained at the end of the course, i.e., women still have a lower sense of belonging and FCI scores than men. This means that the traditional lecture-based course did not help to reduce the gender gaps, and the course outcomes are inequitable as suggested by Rodrigues *et al.*'s equity of parity model [51]. This gender gap may continue to exist unless intentional efforts are made by instructors to create an inclusive and equitable learning environment to help all students develop a high sense of belonging.

We note that even though female students underperformed male students on the FCI, the gender difference in their course grades is relatively small (RQ1), which is consistent with prior studies [77,79,94]. For example, in a prior study, Docktor and Heller reported that even though male students outperformed female students on both pre- and post-FCI tests, there was no gender difference in physics course grades [77]. One possible explanation is that, compared with FCI score, physics course grade consists of several performance measures (e.g., exams, homework, and attendance). In a prior study conducted in an introductory physics course, researchers found that women outperform men on homework and participation, and men outperform women on exams, resulting in course grades of women and men that are not significantly different [171]. Another possible explanation is that the measures of course grade (such as homework and exams) include questions similar to the example questions and practices in class and recitations, while students may not be very familiar with the type of questions in the FCI in traditionally taught courses that primarily emphasize quantitative problem-solving. Prior studies have shown that when people face unfamiliar tasks, they may feel uncertain and anxious [172]. For stereotyped groups in physics, such as women, this uncertainty and anxiety may exacerbate the stereotype threat experienced by them, which has been shown to adversely affect student performance [15,173]. In our study, we find that even though female students have comparable course grades to male students, they still have a lower average sense of belonging than men. This finding is consistent with a prior study showing that women with A grades have the same self-efficacy level as men with C grades [13].

Using structural equation modeling, our study shows that after controlling for students' precollege academic preparation, their sense of belonging in the physics course statistically significantly predicts their FCI scores and course grades (RQ3). If we could eliminate the gender difference in sense of belonging by creating a learning environment in which all students feel valued and feel safe and welcome to engage in collaboration and discussions with peers and instructors, the gender difference in FCI scores and course grades may also decrease. In addition, we find that the strength of the predictive relationship between any two constructs studied is similar for women and men (RQ4). Moreover, we note that even though there are statistically significant gender differences in postsense of belonging, post-FCI scores, and course grades, gender does not directly predict any of them (RQ5), which indicates that the gender differences are mediated through presense of belonging, pre-FCI, and high school GPA. Thus, it is critical to make intentional efforts to reduce these gender gaps keeping in mind Rodrigues *et al.*'s equity of parity model [51].

VIII. IMPLICATIONS

Our study suggests that a sense of belonging is an important predictor of students' physics academic performance. However, we found that in a calculus-based introductory physics course, students' sense of belonging significantly decreased after the course. Moreover, there are persistent gender gaps disadvantaging women in their sense of belonging and academic performance from the beginning to the end of the course. These findings suggest that traditional lecture-based introductory physics courses may not be able to improve students' sense of belonging and reduce the gender gaps unless intentional efforts are made by educators.

As discussed earlier, the weed-out culture in introductory physics courses may contribute to the decrease in students' sense of belonging. Moreover, physics is one of the disciplines with the stereotype about requiring a natural ability to be successful [174,175], and studies have shown that the idea of ability being fixed and unchangeable can further increase students' concerns about belonging, especially for students from underrepresented groups who have few role models [64,176]. Thus, it is critical for instructors to build a learning environment which emphasizes that abilities are malleable and can be changed through deliberate practice and effort [120]. In addition, instructors can also show students nonstereotypical role models from diverse demographic groups, personalities, and interests since this has been shown to increase students' sense of belonging [177,178].

In addition, prior studies show that students' sense of belonging in physics is correlated with whether they feel recognized by instructors and whether they have meaningful and enjoyable interactions with peers [169,179,180]. Therefore, instructors may improve students' sense of belonging by making intentional efforts to positively recognize their students. For example, instructors can recognize students by directly acknowledging their efforts and questions and expressing faith in their ability to excel. In addition to positive recognition, instructors should be careful not to give unintended messages to students, e.g., by praising some students for brilliance or intelligence as opposed to their effort since it can convey to other students that they do not have what is required to excel in physics [181]. Additionally, instructors can promote positive peer interaction, for example, by providing every student with an equal opportunity to express their opinions and avoiding letting a small group of students dominate the discussion.

In addition, there are some research-based classroom interventions that have been shown to enhance students' sense of belonging and reduce gender gaps in students' performance [182–185]. For example, a social belonging intervention, which focuses on establishing a classroom climate in which adversity is framed as a normal student experience and likely to be overcome by working hard and working smart and taking advantage of all of the resources, has been shown to reduce the gender difference in students' physics course grades [182]. Instructors can tailor the research-based interventions in their classes to help all students develop a higher sense of belonging.

IX. LIMITATIONS AND FUTURE DIRECTIONS

In this study, we investigated students' sense of belonging, physics conceptual understanding, and course grades in an introductory calculus-based physics course. We used Force Concept Inventory to measure students' conceptual understanding, which is commonly used in physics education research but mainly focuses on force concepts and kinematics [179,186–188]. Another widely used concept inventory focusing on Newtonian mechanics is force and motion conceptual evaluation (FMCE) [80]. A prior study shows that students' scores on the FCI and FMCE are strongly related, while these two inventories have slightly different content domains and representational formats [125]. For example, the FMCE is designed to measure student understanding of one-dimensional force and motion, while FCI has a broader domain that also includes two-dimensional motion and a wider application of forces in more diverse settings [125]. In future study, it would be helpful to conduct similar studies using FMCE or other concept inventories focusing on Newtonian mechanics to examine whether findings are similar. It would also be helpful to conduct similar studies in the calculus-based introductory electricity and magnetism using the Conceptual Survey of Electricity and Magnetism) to measure students' conceptual understanding.

In addition, this study is based on students' self-reported responses to the survey. It would be helpful to interview more students to get a deeper understanding of the mechanism of how students' sense of belonging is shaped by their learning experience in the physics course and how their sense of belonging influences their academic outcomes. It would also be helpful to use belonging interventions with control groups to further study the effect of sense of belonging on students' conceptual understanding and course grades.

This study was conducted in a primarily traditionally taught introductory calculus-based physics course. It would be interesting to investigate students' sense of belonging in courses with different class formats and teaching approaches, such as active engagement pedagogies. It would also be valuable to conduct similar studies in the classes in which there is an intentional effort in promoting students' sense of belonging and compare the results with those of the current study. Future studies can also investigate students' sense of belonging and conceptual understanding in other physics courses, such as advanced physics courses beyond the first year, which are typically taken by physics majors. In addition, our study was conducted in a large public research university in the United States. Similar studies in different types of institutions such as small colleges and universities in the United States and in other countries would also be helpful for developing a deeper understanding of the relationships between students' sense of belonging and their physics conceptual understanding and course grades.

ACKNOWLEDGMENTS

This work was supported by Grant No. DUE-1524575 from the National Science Foundation. We thank Dr. Robert P. Devaty for his constructive feedback on the manuscript.

APPENDIX: SEM MODEL WITH ALL STUDENTS' DATA

In the results section of the main text, we presented the SEM model (Fig. 3) using matched students from pre to post. Here, we present the SEM model (Fig. 4) with all students' data (unmatched from pre to post). Comparing the two models, we note that the models are very similar, which suggests that the unmatched students' data do not have systematic missing data.

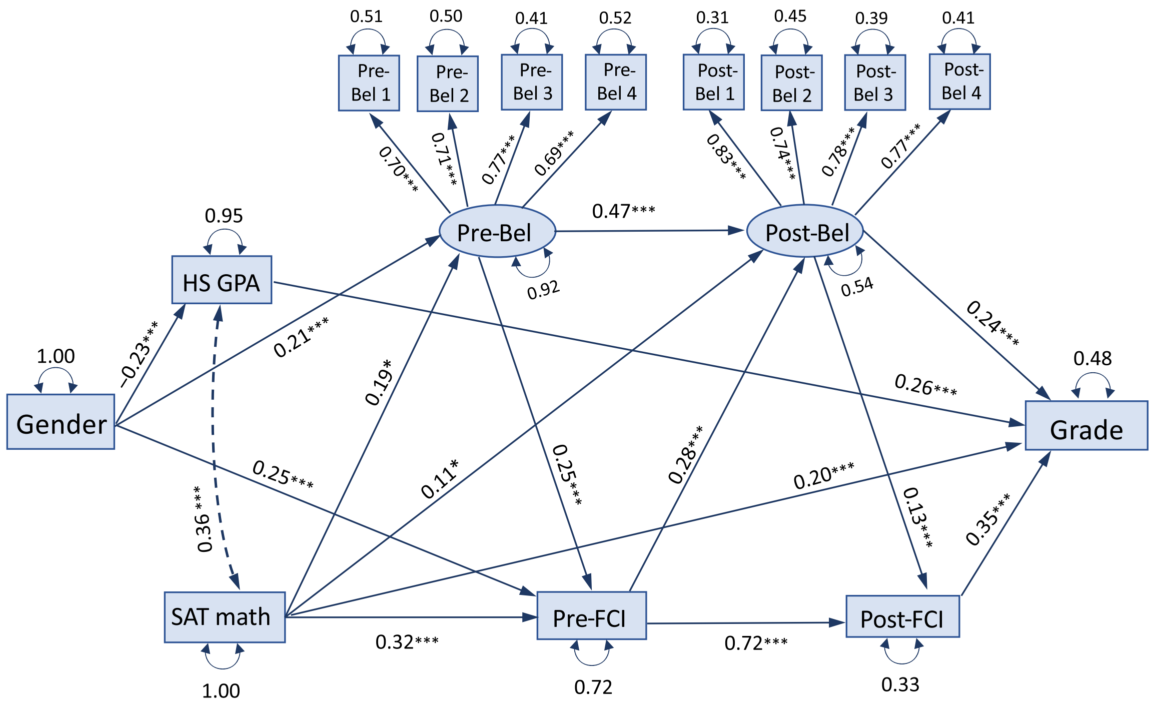

FIG. 4. Results of the SEM model with all students' data (unmatched from pre to post).

[1] H. B. Carlone and A. Johnson, Understanding the science experiences of successful women of color: Science identity as an analytic lens, J. Res. Sci. Teach. **44**, 1187 (2007).
[2] 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).
[3] Z. Hazari, G. Sonnert, P. M. Sadler, and M.-C. Shanahan, Connecting high school physics experiences, outcome expectations, physics identity, and physics career choice: A gender study, J. Res. Sci. Teach. **47**, 978 (2010).
[4] J. M. Nissen and J. T. Shemwell, Gender, experience, and self-efficacy in introductory physics, Phys. Rev. Phys. Educ. Res. **12**, 020105 (2016).
[5] V. Sawtelle, E. Brewe, and L. H. Kramer, Exploring the relationship between self-efficacy and retention in introductory physics, J. Res. Sci. Teach. **49**, 1096 (2012).
[6] B. Van Dusen and J. Nissen, Equity in college physics student learning: A critical quantitative intersectionality investigation, J. Res. Sci. Teach. **57**, 33 (2020).
[7] M. Lorenzo, C. H. Crouch, and E. Mazur, Reducing the gender gap in the physics classroom, Am. J. Phys. **74**, 118 (2006).
[8] K. Rosa and F. M. Mensah, Educational pathways of Black women physicists: Stories of experiencing and overcoming obstacles in life, Phys. Rev. Phys. Educ. Res. **12**, 020113 (2016).
[9] 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).
[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] I. Rodriguez, G. Potvin, and L. H. Kramer, How gender and reformed introductory physics impacts student success in advanced physics courses and continuation in the physics major, Phys. Rev. Phys. Educ. Res. **12**, 020118 (2016).

[12] 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).

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

[14] C. M. Steele and J. Aronson, Stereotype threat and the intellectual test performance of African Americans, J. Pers. Soc. Psychol. **69**, 797 (1995).

[15] G. C. Marchand and G. Taasoobshirazi, Stereotype threat and women's performance in physics, Int. J. Sci. Educ. **35**, 3050 (2013).

[16] R. Ivie, S. White, and R. Y. Chu, Women's and men's career choices in astronomy and astrophysics, Phys. Rev. Phys. Educ. Res. **12**, 020109 (2016).

[17] 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).

[18] 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).

[19] J. Blue, A. L. Traxler, and X. C. Cid, Gender matters, Phys. Today **71**, No. 3 40 (2018).

[20] S. J. Pollock, N. D. Finkelstein, and L. E. Kost, Reducing the gender gap in the physics classroom: How sufficient is interactive engagement? Phys. Rev. ST Phys. Educ. Res. **3**, 010107 (2007).

[21] E. Brewe, V. Sawtelle, L. H. Kramer, G. E. O'Brien, I. Rodriguez, and P. Pamelá, Toward equity through participation in modeling instruction in introductory university physics, Phys. Rev. ST Phys. Educ. Res. **6**, 010106 (2010).

[22] G. Sonnert and M. F. Fox, Women, men, and academic performance in science and engineering: The gender difference in undergraduate grade point averages, J. Higher Educ. **83**, 73 (2012).

[23] 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).

[24] A. Malespina and C. Singh, Impact of grade penalty in first-year foundational science courses on female engineering majors, Int. J. Eng. Educ. **38**, 1021 (2022), https://www.ijee.ie/1atestissues/Vol38-4/13_ijee4218.pdf.

[25] M.-T. Wang and J. Degol, Motivational pathways to STEM career choices: Using expectancy-value perspective to understand individual and gender differences in STEM fields, Dev. Rev. **33**, 304 (2013).

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

[27] J. S. Eccles, Understanding women's educational, and occupational choices: Applying the Eccles, *et al.* model of achievement-related choices, Psychol. Women Q. **18**, 585 (1994).

[28] J. L. Smith, C. Sansone, and P. H. White, The stereotyped task engagement process: The role of interest and achievement motivation, J. Educ. Psychol. **99**, 99 (2007).

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

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

[31] A. L. Zeldin, S. L. 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).

[32] N. E. Betz and G. Hackett, Applications of self-efficacy theory to understanding career choice behavior, J. Soc. Clin. Psychol. **4**, 279 (1986).

[33] E. Lichtenberger and C. George-Jackson, Predicting high school students' interest in majoring in a STEM field: Insight into high school students' postsecondary plans, J. Career Tech. Educ. **28**, 19 (2013).

[34] V. Tinto, Classrooms as communities: Exploring the educational character of student persistence, J. Higher Educ. **68**, 599 (1997).

[35] J. C. Blickenstaff, Women and science careers: Leaky pipeline or gender filter? Gender Educ. **17**, 369 (2005).

[36] S. Cwik and C. Singh, Self-efficacy and perceived recognition by peers, instructors, and teaching assistants in physics predict bioscience majors' science identity, PLoS One **17**, e0273621 (2022).

[37] A. Malespina, C. D. Schunn, and C. Singh, Whose ability and growth matter? Gender, mindset and performance in physics, Int. J. STEM Educ. **9**, 28 (2022).

[38] Y. Li and C. Singh, How engineering identity of first-year female and male engineering majors is predicted by their physics self-efficacy and identity, Int. J. Eng. Educ. **38**, 799 (2022), https://www.ijee.ie/latestissues/Vol38-3/21_ijee4203.pdf.

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

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

[41] S. J. Correll, Gender and the career choice process: The role of biased self-assessments, Am. J. Sociol. **106**, 1691 (2001).

[42] S. J. Correll, Constraints into preferences: Gender, status, and emerging career aspirations, Am. Sociol. Rev. **69**, 93 (2004).

[43] R. Elliott, A. C. Strenta, R. Adair, M. Matier, and J. Scott, The role of ethnicity in choosing and leaving science in highly selective institutions, Res. High. Educ. **37**, 681 (1996).

[44] E. T. Pascarella and P. T. Terenzini, *How College Affects Students: Findings and Insights from Twenty Years of Research* (Jossey-Bass, San Francisco, CA, 1991).

[45] C. Hill, C. Corbett, and A. St. Rose, *Why so Few? Women in Science, Technology, Engineering, and Mathematics* (American Association of University Women, Washington, DC, 2010).

[46] S. Ahlqvist, B. London, and L. Rosenthal, Unstable identity compatibility: How gender rejection sensitivity undermines the success of women in science, technology, engineering, and mathematics fields, Psychol. Sci. **24**, 1644 (2013).

[47] A. Christopher Strenta, R. Elliott, R. Adair, M. Matier, and J. Scott, Choosing and leaving science in highly selective institutions, Res. High. Educ. **35**, 513 (1994).

[48] A. B. Diekman, E. K. Clark, A. M. Johnston, E. R. Brown, and M. Steinberg, Malleability in communal goals and beliefs influences attraction to STEM careers: Evidence for a goal congruity perspective, J. Pers. Soc. Psychol. **101**, 902 (2011).

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

[50] 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).

[51] 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).

[52] G. M. Walton and G. L. Cohen, A question of belonging: Race, social fit, and achievement, J. Pers. Soc. Psychol. **92**, 82 (2007).

[53] K. L. Lewis, J. G. Stout, N. D. Finkelstein, S. J. Pollock, A. Miyake, G. L. Cohen, and T. A. Ito, Fitting in to move forward: Belonging, gender, and persistence in the physical sciences, technology, engineering, and mathematics (pSTEM), Psychol. Women Q. **41**, 420 (2017).

[54] A. H. Maslow, A dynamic theory of human motivation, in *Understanding Human Motivation* (1958).

[55] K. L. Lewis, J. G. Stout, S. J. Pollock, N. D. Finkelstein, and T. A. Ito, Fitting in or opting out: A review of key social-psychological factors influencing a sense of belonging for women in physics, Phys. Rev. Phys. Educ. Res. **12**, 020110 (2016).

[56] L. R. Hausmann, J. W. Schofield, and R. L. Woods, Sense of belonging as a predictor of intentions to persist among African American and White first-year college students, Res. High. Educ. **48**, 803 (2007).

[57] L. D. Pittman and A. Richmond, Academic and psychological functioning in late adolescence: The importance of school belonging, J. Exp. Educ. **75**, 270 (2007).

[58] 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).

[59] 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).

[60] A. J. Kola, Investigating the conceptual understanding of physics through an interactive lecture-engagement, Cumhuriyet Int. J. Educ. **6**, 82 (2017).

[61] A. Madsen, S. B. McKagan, and E. C. Sayre, Best practices for administering concept inventories, Phys. Teach. **55**, 530 (2017).

[62] V. Seyranian, A. Madva, N. Duong, N. Abramzon, Y. Tibbetts, and J. M. Harackiewicz, The longitudinal effects of STEM identity and gender on flourishing and achievement in college physics, Int. J. STEM Educ. **5**, 40 (2018).

[63] A. Ladewig, M. Keller, and U. Klusmann, Sense of belonging as an important factor in the pursuit of physics: Does it also matter for female participants of the German physics olympiad? Front. Psychol. **11**, 548781 (2020).

[64] A. Deiglmayr, E. Stern, and R. Schubert, Beliefs in "brilliance" and belonging uncertainty in male and female STEM students, Front. Psychol. **10**, 1114 (2019).

[65] K. A. Bailey, D. Horacek, S. Worthington, A. Nanthakumar, S. Preston, and C. C. Ilie, STEM/non-STEM divide structures undergraduate beliefs about gender and talent in academia, Front. Sociol. **4**, 26 (2019).

[66] S. Cwik and C. Singh, Framework for and review of research on assessing and improving equity and inclusion in undergraduate physics learning environments, in *International Handbook of Physics Education Research: Special Topics*, edited by M. F. Taşar and P R. L. Heron (AIP Publishing, Melville, NY, 2023), https://doi.org/10.1063/9780735425514_002.

[67] S. Cwik and C. Singh, Not feeling recognized as a physics person by instructors and teaching assistants is correlated with female students' lower grades, Phys. Rev. Phys. Educ. Res. **18**, 010138 (2022).

[68] S. Cwik and C. Singh, How perception of learning environment predicts male and female students' grades and motivational outcomes in algebra-based introductory physics courses, Phys. Rev. Phys. Educ. Res. **17**, 020143 (2021).

[69] S. Cwik and C. Singh, Developing an innovative sustainable science education ecosystem: Lessons from negative impacts of inequitable and non-inclusive learning environments, Sustainability **14**, 11345 (2022).

[70] Y. Li and C. Singh, The impact of perceived recognition by physics instructors on women's self-efficacy and interest, arXiv.2303.07239.

[71] L. M. Santana and C. Singh, Negative impacts of an unwelcoming physics environment on undergraduate women, in *Proceedings of the Physics Education Research Conference (PERC)* (2021), p. 377, 10.1119/perc.2021.pr.Santana.

[72] D. Doucette, R. Clark, and C. Singh, Hermione and the Secretary: How gendered task division in introductory physics labs can disrupt equitable learning, Eur. J. Phys. **41**, 035702 (2020).

[73] D. Doucette and C. Singh, Why are there so few women in physics? Reflections on the experiences of two women, Phys. Teach. **58**, 297 (2020).

[74] K. Lewis, J. Stout, N. Finkelstein, S. Pollock, A. Miyake, G. Cohen, and T. Ito, Fitting in to move forward: Using a belonging framework to understand gender disparities in persistence in the physical sciences, technology, engineering and mathematics (pSTEM), Psychol. Women Q. **41**, 420 (2017).

[75] L. McCullough, Gender differences in student responses to physics conceptual questions based on question context, *Proceedings of the Physics Education Research Conference* (ASQ Advancing the STEM Agenda in Education, 2011).

[76] R. R. Hake, Relationship of individual student normalized learning gains in mechanics with gender, high-school physics, and pretest scores on mathematics and spatial visualization, in *Proceedings of the 2002 Physics Education Research Conference, Boise, Idaho* (AIP, New York, 2002), p. 1.

[77] J. Docktor and K. Heller, Gender differences in both force concept inventory and introductory physics performance, in *AIP Conference Proceedings* (American Institute of Physics, 2008), p. 15.

[78] C. T. Richardson and B. W. O'Shea, Assessing gender differences in response system questions for an introductory physics course, Am. J. Phys. **81**, 231 (2013).

[79] S. Bates, R. Donnelly, C. MacPhee, D. Sands, M. Birch, and N. R. Walet, Gender differences in conceptual understanding of Newtonian mechanics: A UK cross-institution comparison, Eur. J. Phys. **34**, 421 (2013).

[80] R. K. Thornton and D. R. Sokoloff, Assessing student learning of Newton's laws: The force and motion conceptual evaluation and the evaluation of active learning laboratory and lecture curricula, Am. J. Phys. **66**, 338 (1998).

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

[82] L. McCullough, Gender, context, and physics assessment, J. Int. Women's Stud. **5**, 20 (2004), https://vc.bridgew.edu/jiws/vol5/iss4/2.

[83] R. Henderson, P. Miller, J. Stewart, A. Traxler, and R. Lindell, Item-level gender fairness in the force and motion conceptual evaluation and the conceptual survey of electricity and magnetism, Phys. Rev. Phys. Educ. Res. **14**, 020103 (2018).

[84] A. Traxler, R. Henderson, J. Stewart, G. Stewart, A. Papak, and R. Lindell, Gender fairness within the Force Concept Inventory, Phys. Rev. Phys. Educ. Res. **14**, 010103 (2018).

[85] R. Henderson, J. Stewart, and A. Traxler, Partitioning the gender gap in physics conceptual inventories: Force concept inventory, force and motion conceptual evaluation, and conceptual survey of electricity and magnetism, Phys. Rev. Phys. Educ. Res. **15**, 010131 (2019).

[86] V. P. Coletta, J. A. Phillips, and J. Steinert, FCI normalized gain, scientific reasoning ability, thinking in physics, and gender effects, in *AIP Conference Proceedings* (American Institute of Physics, 2012), p. 23.

[87] V. P. Coletta and J. A. Phillips, Interpreting FCI scores: Normalized gain, preinstruction scores, and scientific reasoning ability, Am. J. Phys. **73**, 1172 (2005).

[88] T. L. McCaskey, M. H. Dancy, and A. Elby, Effects on assessment caused by splits between belief and understanding, in *AIP Conference Proceedings* (American Institute of Physics, 2004), p. 37.

[89] T. L. McCaskey and A. Elby, Probing students' epistemologies using split tasks, in *AIP Conference Proceedings* (American Institute of Physics, 2005), p. 57.

[90] 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).

[91] P. B. Kohl and H. V. Kuo, Introductory physics gender gaps: Pre-and post-studio transition, in *AIP Conference Proceedings* (American Institute of Physics, 2009), p. 173.

[92] M. J. Cahill, K. M. Hynes, R. Trousil, L. A. Brooks, M. A. McDaniel, M. Repice, J. Zhao, and R. F. Frey, Multiyear, multi-instructor evaluation of a large-class interactive-engagement curriculum, Phys. Rev. ST Phys. Educ. Res. **10**, 020101 (2014).

[93] A. Maries, N. Karim, and C. Singh, Active learning in an inequitable learning environment can increase the gender performance gap: The negative impact of stereotype threat, Phys. Teach. **58**, 430 (2020).

[94] R. Henderson, G. Stewart, J. Stewart, L. Michaluk, and A. Traxler, Exploring the gender gap in the conceptual survey of electricity and magnetism, Phys. Rev. Phys. Educ. Res. **13**, 020114 (2017).

[95] 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).

[96] C. Zabriskie, J. Yang, S. DeVore, and J. Stewart, Using machine learning to predict physics course outcomes, Phys. Rev. Phys. Educ. Res. **15**, 020120 (2019).

[97] Z. Zhu, Gender differences in mathematical problem solving patterns: A review of literature, Int. Educ. J. **8**, 187 (2007), https://api.semanticscholar.org/CorpusID:6079139.

[98] B. Bridgeman and C. Lewis, Gender differences in college mathematics grades and SAT-M scores: A reanalysis of Wainer and Steinberg, J. Educ. Measure. **33**, 257 (1996).

[99] D. Reilly, Gender, culture, and sex-typed cognitive abilities, PLoS One **7**, e39904 (2012).

[100] D. M. Quinn and S. J. Spencer, The interference of stereotype threat with women's generation of mathematical problem-solving strategies, J. Soc. Issues **57**, 55 (2001).

[101] M. Niederle and L. Vesterlund, Explaining the gender gap in math test scores: The role of competition, J. Econ. Perspect. **24**, 129 (2010).

[102] C. M. Steele, Stereotyping and its threat are real, Am. Psychol. **53**, 680 (1998).

[103] 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).

[104] M. Walsh, C. Hickey, and J. Duffy, Influence of item content and stereotype situation on gender differences in mathematical problem solving, Sex Roles **41**, 219 (1999).

[105] N. M. Fortin, P. Oreopoulos, and S. Phipps, Leaving boys behind gender disparities in high academic achievement, J. Health Hum. Resour. Adm. **50**, 549 (2015).

[106] B. Bridgeman and C. Wendler, Gender differences in predictors of college mathematics performance and in

college mathematics course grades, J. Educ. Psychol. **83**, 275 (1991).
[107] J. E. Rosenbaum, *Beyond College for all: Career Paths for the Forgotten Half* (Russell Sage Foundation, New York, NY, 2001).
[108] S. Kelly, What types of students' effort are rewarded with high marks?, Sociol. Educ. **81**, 32 (2008).
[109] G. Farkas, R. P. Grobe, D. Sheehan, and Y. Shuan, Cultural resources and school success: Gender, ethnicity, and poverty groups within an urban school district, Am. Sociol. Rev. **55**, 127 (1990).
[110] S. Klevan, S. L. Weinberg, and J. A. Middleton, Why the boys are missing: Using social capital to explain gender differences in college enrollment for public high school students, Res. High. Educ. **57**, 223 (2016).
[111] T. A. DiPrete and C. Buchmann, *The Rise of Women: The Growing Gender Gap in Education And What It Means for American Schools* (Russell Sage Foundation, New York, NY, 2013).
[112] L. Sparkman, W. Maulding, and J. Roberts, Non-cognitive predictors of student success in college, Coll. Stud. J. **46**, 642 (2012), https://api.semanticscholar.org/CorpusID:140406099.
[113] Q. X. Ryan, D. Del Agunos, A. Villasenor, H. R. Sadaghiani, and A. Small, Evaluating students' performance on the FCI at a minority serving institution, presented at PER Conf. 2019, Provo, UT, 10.1119/perc.2019.pr.Ryan.
[114] V. P. Coletta, J. A. Phillips, and J. J. Steinert, Interpreting Force Concept Inventory scores: Normalized gain and SAT scores, Phys. Rev. ST Phys. Educ. Res. **3**, 010106 (2007).
[115] J. D. Edwards, R. S. Barthelemy, and R. F. Frey, Relationship between course-level social belonging (sense of belonging and belonging uncertainty) and academic performance in General Chemistry 1, J. Chem. Educ. **99**, 71 (2021).
[116] G. L. Cohen and J. Garcia, Identity, belonging, and achievement: A model, interventions, implications, Curr. Dir. Psychol. Sci. **17**, 365 (2008).
[117] D. Hewagallage, E. Christman, and J. Stewart, Examining the relation of high school preparation and college achievement to conceptual understanding, Phys. Rev. Phys. Educ. Res. **18**, 010149 (2022).
[118] L. J. Sax, J. M. Blaney, K. J. Lehman, S. L. Rodriguez, K. L. George, and C. Zavala, Sense of belonging in computing: The role of introductory courses for women and underrepresented minority students, Soc. Sci. **7**, 122 (2018).
[119] S. Krause-Levy, W. G. Griswold, L. Porter, and C. Alvarado, The relationship between sense of belonging and student outcomes in CS1 and beyond, in *Proceedings of the 17th ACM Conference on International Computing Education Research* (Association for Computing Machinery, New York, NY, 2021), p. 29.
[120] C. Good, A. Rattan, and C. S. Dweck, Why do women opt out? Sense of belonging and women's representation in mathematics, J. Pers. Soc. Psychol. **102**, 700 (2012).
[121] J. J. Hox and T. M. Bechger, An introduction to structural equation modeling, Family Sci. Rev. **11**, 354 (1998).
[122] L. Ding, Theoretical perspectives of quantitative physics education research, Phys. Rev. Phys. Educ. Res. **15**, 020101 (2019).
[123] S. M. Stoen, M. A. McDaniel, R. F. Frey, K. M. Hynes, and M. J. Cahill, Force Concept Inventory: More than just conceptual understanding, Phys. Rev. Phys. Educ. Res. **16**, 010105 (2020).
[124] I. A. Halloun and D. Hestenes, The initial knowledge state of college physics students, Am. J. Phys. **53**, 1043 (1985).
[125] R. K. Thornton, D. Kuhl, K. Cummings, and J. Marx, Comparing the force and motion conceptual evaluation and the force concept inventory, Phys. Rev. ST Phys. Educ. Res. **5**, 010105 (2009).
[126] 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).
[127] N. Lasry, S. Rosenfield, H. Dedic, A. Dahan, and O. Reshef, The puzzling reliability of the Force Concept Inventory, Am. J. Phys. **79**, 909 (2011).
[128] J. Stewart, H. Griffin, and G. Stewart, Context sensitivity in the force concept inventory, Phys. Rev. ST Phys. Educ. Res. **3**, 010102 (2007).
[129] D. Hammer, Epistemological beliefs in introductory physics, Cognit. Instr. **12**, 151 (1994).
[130] Learning Activation Lab, Activation lab tools: Measures, and data collection instruments (2017), http://www.activationlab.org/tools/ (Accessed February 4, 2019).
[131] J. Schell and B. Lukoff, Peer instruction self-efficacy instrument [Developed at Harvard University] (unpublished) (2010).
[132] PERTS Academic Mindsets Assessment (2020), https://www.perts.net/orientation/ascend (Accessed February 3, 2019).
[133] Z. Y. Kalender, E. Marshman, T. J. Nokes-Malach, C. D. Schunn, and C. Singh, Motivational characteristics of underrepresented ethnic and racial minority students in introductory physics courses, presented at PER Conf. 2017, Cincinnati, OH, 10.1119/perc.2017.pr.046.
[134] T. 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, presented at PER Conf. 2017, Cincinnati, OH, 10.1119/perc.2017.pr.064.
[135] Z. Y. Kalender, E. Marshman, C. D. Schunn, T. J. Nokes-Malach, and C. Singh, Gendered patterns in the construction of physics identity from motivational factors, Phys. Rev. Phys. Educ. Res. **15**, 020119 (2019).
[136] Z. Y. Kalender, E. Marshman, C. D. Schunn, T. J. Nokes-Malach, and C. Singh, Large gender differences in physics self-efficacy at equal performance levels: A warning sign?, presented at PER Conf. 2018, Washington, DC, 10.1119/perc.2018.pr.Kalender.
[137] B. Thompson, *Exploratory and Confirmatory Factor Analysis* (American Psychological Association, Washington, 2004).
[138] K. Pearson and F. Galton, VII. Note on regression and inheritance in the case of two parents, Proc. R. Soc. London **58**, 240 (1895).

[139] L. J. Cronbach, Coefficient alpha and the internal structure of tests, Psychometrika **16**, 297 (1951).
[140] C. Goodenow, Classroom belonging among early adolescent students: Relationships to motivation and achievement, J. Early Adolesc. **13**, 21 (1993).
[141] J. F. Hair, W. C. Black, B. J. Babin, and R. E. Anderson, *Multivariate Data Analysis: International Version* (Pearson, New Jersey, 2010).
[142] Z. Y. Kalender, E. Marshman, C. D. Schunn, T. J. 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**, 020148 (2019).
[143] S. E. Embretson and S. P. Reise, *Item Response Theory for Psychologists* (Lawrence Erlbaum Associates Publishers, Mahwah, NJ, 2000).
[144] https://depts.washington.edu/uwpeg/research/publications.
[145] R. P. Chalmers, mirt: A multidimensional item response theory package for the R environment, J. Stat. Softw. **48**, 1 (2012).
[146] F. Samejima, Estimation of latent ability using a response pattern of graded scores, in *Psychometrika Monograph* (Psychometric Society, Richmond, VA, 1969), p. 17.
[147] A. V. Knaub, J. M. Aiken, and L. Ding, Two-phase study examining perspectives and use of quantitative methods in physics education research, Phys. Rev. Phys. Educ. Res. **15**, 020102 (2019).
[148] R. P. Springuel, M. C. Wittmann, and J. R. Thompson, Reconsidering the encoding of data in physics education research, Phys. Rev. Phys. Educ. Res. **15**, 020103 (2019).
[149] J. Pallant, *SPSS Survival Manual: A Step by Step Guide to Data Analysis Using IBM SPSS* (Routledge, London, 2020).
[150] R Core Team, *R: A Language and Environment for Statistical Computing* (R Foundation, 2013), https://www.r-project.org/ (Accessed 6 February 2019).
[151] A. J. Tomarken and N. G. Waller, Structural equation modeling: Strengths, limitations, and misconceptions, Annu. Rev. Clin. Psychol. **1**, 31 (2005).
[152] R. B. Kline, *Principles and Practice of Structural Equation Modeling* (Guilford Publications, New York, NY, 2015).
[153] D. Kaplan, *Structural Equation Modeling: Foundations and Extensions* (Sage Publications, 2008), Vol. 10.
[154] D. Kaplan, Structural equation modeling, in *International Encyclopedia of the Social, and Behavioral Sciences*, edited by N. J. Smelser and P. B. Baltes (Pergamon Press, Oxford, 2001), p. 15215.
[155] Z. Awang, *SEM Made Simple: A Gentle Approach to Learning Structural Equation Modeling* (MPWS Rich Publication, Bandar Baru Bangi, 2015).
[156] A. G. Yong and S. Pearce, A beginner's guide to factor analysis: Focusing on exploratory factor analysis, Tutorials Quant. Methods Psychol. **9**, 79 (2013).
[157] D. Mindrila, Maximum likelihood (ML) and diagonally weighted least squares (DWLS) estimation procedures: A comparison of estimation bias with ordinal and multivariate non-normal data, Int. J. Digital Soc. **1**, 60 (2010).
[158] M. Rhemtulla, P. É. Brosseau-Liard, and V. Savalei, When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions, Psychol. Methods **17**, 354 (2012).
[159] D. Hooper, J. Coughlan, and M. Mullen, Structural equation modeling: Guidelines for determining model fit, Electron. J. Bus. Res. Methods **6**, 53 (2007).
[160] K. S. Betts, G. M. Williams, J. M. Najman, and R. Alati, The role of sleep disturbance in the relationship between post-traumatic stress disorder and suicidal ideation, Journal of anxiety disorders **27**, 735 (2013).
[161] C. A. Kronauge, The effects of mixing metrics and distributions simultaneously in structural equation modeling: A simulation study, Doctoral dissertation, University of Northern Colorado, 2012.
[162] G. D. Garson, *Structural Equation Modeling* (G. David Garson and Statistical Publishing Associates, 2014).
[163] P. W. Lei and Q. Wu, Introduction to structural equation modeling: Issues and practical considerations, Educ. Meas. **26**, 33 (2007).
[164] R. P. Bagozzi and Y. Yi, On the evaluation of structural equation models, J. Acad. Market. Sci. **16**, 74 (1988).
[165] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences* (L. Erlbaum Associates, Hillsdale, NJ, 1988).
[166] K. Siddiqui, Heuristics for sample size determination in multivariate statistical techniques, World Appl. Sci. J. **27**, 285 (2013).
[167] D. F. Halpern, Validity, fairness, and group differences: Tough questions for selection testing, Psychol. Publ. Pol. Law, **6**, 56 (2000).
[168] E. S. Spelke, Sex differences in intrinsic aptitude for mathematics and science?: A critical review, Am. Psychol. **60**, 950 (2005).
[169] 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).
[170] Y. Li, Improving inclusiveness of learning environment in introductory physics and content understanding in quantum physics, Doctoral dissertation, University of Pittsburgh, 2022, http://d-scholarship.pitt.edu/id/eprint/43434.
[171] L. E. Kost-Smith, S. J. Pollock, and N. D. Finkelstein, Gender disparities in second-semester college physics: The incremental effects of a "smog of bias", Phys. Rev. ST Phys. Educ. Res. **6**, 020112 (2010).
[172] B. B. Hudson, Anxiety in response to the unfamiliar, J. Soc. Issues **10**, 53 (1954).
[173] A. Maries, N. I. Karim, and C. Singh, Does stereotype threat affect female students' performance in introductory physics? AIP Conf. Proc. **2109**, 120001 (2019).
[174] L. Bian, S.-J. Leslie, and A. Cimpian, Gender stereotypes about intellectual ability emerge early and influence children's interests, Science **355**, 389 (2017).
[175] S.-J. Leslie, A. Cimpian, M. Meyer, and E. Freeland, Expectations of brilliance underlie gender distributions across academic disciplines, Science **347**, 262 (2015).
[176] L. Bian, S.-J. Leslie, M. C. Murphy, and A. Cimpian, Messages about brilliance undermine women's interest in educational and professional opportunities, J. Exp. Soc. Psychol. **76**, 404 (2018).

[177] S. Cheryan, B. J. Drury, and M. Vichayapai, Enduring influence of stereotypical computer science role models on women's academic aspirations, Psychol. Women Q. **37**, 72 (2013).

[178] J. G. Stout, N. Dasgupta, M. Hunsinger, and M. A. McManus, STEMing the tide: Using ingroup experts to inoculate women's self-concept in science, technology, engineering, and mathematics (STEM), J. Pers. Soc. Psychol. **100**, 255 (2011).

[179] Z. Hazari, D. Chari, G. Potvin, and E. Brewe, The context dependence of physics identity: Examining the role of performance/competence, recognition, interest, and sense of belonging for lower and upper female physics undergraduates, J. Res. Sci. Teach. **57**, 1583 (2020).

[180] Y. Li and C. Singh, Inclusive learning environments can improve student learning and motivational beliefs, Phys. Rev. Phys. Educ. Res. **18**, 020147 (2022).

[181] Y. Li, K. Whitcomb, and C. Singh, How perception of being recognized or not recognized by instructors as a "physics person" impacts male and female students' self-efficacy and performance, Phys. Teach. **58**, 484 (2020).

[182] K. R. Binning, N. Kaufmann, E. M. McGreevy, O. Fotuhi, S. Chen, E. Marshman, Z. Y. Kalender, L. Limeri, L. Betancur, and C. Singh, Changing social contexts to foster equity in college science courses: An ecological-belonging intervention, Psychol. Sci. **31**, 1059 (2020).

[183] A. Master and A. N. Meltzoff, Cultural stereotypes and sense of belonging contribute to gender gaps in STEM, Int. J. Gender Sci. Technol. **12**, 152 (2020), https://genderandset.open.ac.uk/index.php/genderandset/article/view/674.

[184] D. S. Yeager and G. M. Walton, Social-psychological interventions in education: They're not magic, Rev. Educ. Res. **81**, 267 (2011).

[185] 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, relationships, and achievement in engineering, J. Educ. Psychol. **107**, 468 (2015).

[186] 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).

[187] A. Godwin, G. Potvin, Z. Hazari, and R. Lock, Understanding engineering identity through structural equation modeling, in *Proceedings of 2013 IEEE Frontiers in Education Conference (FIE), Oklahoma City, OK* (IEEE, New York, 2013), p. 50.

[188] R. M. Lock, Z. Hazari, and G. Potvin, Impact of out-of-class science and engineering activities on physics identity and career intentions, Phys. Rev. Phys. Educ. Res. **15**, 020137 (2019).