Inclusive learning environments can improve student learning and motivational beliefs

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

(Received 25 January 2022; accepted 24 October 2022; published 23 December 2022)

We discuss an investigation of students' motivational beliefs and performance on the Force Concept Inventory (FCI) in a calculus-based introductory physics course at a large public university in the U.S. We investigated how students' perception of the inclusiveness of the learning environment (including perceived recognition, perceived effectiveness of peer interaction, and sense of belonging) predicts students' FCI scores and physics motivational beliefs (including self-efficacy, interest, and overall physics identity) at the end of the course after controlling for students' high school performance and their FCI scores and motivational beliefs at the beginning of the course. We find signatures of noninclusive learning environment in that female students' mean scores in physics motivational beliefs and perception of the inclusiveness of the learning environment were lower than male students', and the gender gap in students' self-efficacy increased from the beginning to the end of the course. Using structural equation modeling, we find that the gender differences in students' motivational beliefs and FCI scores were mediated by the different components of students' perception of the inclusiveness of the learning environment. In particular, students' perceived recognition, e.g., by instructors, was an important predictor of their overall physics identity, and their sense of belonging predicted their self-efficacy and FCI scores. Our findings can be valuable for contemplating guidelines for creating an inclusive learning environment in which all students can excel.

DOI: 10.1103/PhysRevPhysEducRes.18.020147

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 U.S., only 20% of the physics undergraduate degrees are earned by women [19]. In addition, several studies have also reported gender disparity in students' performance in some STEM disciplines [20–22]. Prior research suggests that individuals' performance and persistence in STEM can be influenced by their motivational beliefs such as self-efficacy, interest, and identity in that domain [1–3,5,10,18,23–35]. Students from underrepresented groups in STEM such as women may not have enough encouragement and role models to help them develop strong motivational beliefs in STEM. In addition, the societal stereotypes and biases in STEM may further undermine their motivational beliefs and could lead to withdrawal from STEM courses, majors, or careers [36–46]. Therefore, investigation of the factors that can influence students' motivational beliefs and performance is important to understanding the underrepresentation, e.g., of women and other marginalized students in STEM, and can help in developing guidelines for building an inclusive learning environment and promoting diversity and equity in STEM fields.

By inclusive learning environment, we refer to an environment in which all students feel welcome, valued, and supported. By equity in learning, we mean that not only should all students have adequate opportunities and access to resources, and have an inclusive learning environment with appropriate support and mentoring so that they can engage in learning in a meaningful and enjoyable manner, but the course outcomes should be equitable. Therefore, inclusiveness is necessary but not sufficient for equity since inclusiveness does not guarantee equitable course outcomes. By equitable course outcomes, we mean that students from all demographic groups (e.g., regardless of their gender identity or race or ethnicity) who have the prerequisites to enroll in the course, on average, have comparable outcomes, which is consistent with Rodriguez *et al.*'s equity of parity model [47]. The STEM course outcomes include student performance and their STEM motivational beliefs at the end of the courses because regardless of the performance, the motivational beliefs can influence students' short-term and long-term retention in STEM disciplines [23,48]. We note that adequate opportunity and access to resources, inclusive learning environment, and equitable outcomes are strongly entangled with each other. For example, if the learning environment is not inclusive, the outcomes are unlikely to be equitable. In this study, we aim to understand how students' perception of the inclusiveness of the learning environment predicts their course outcomes including both academic performance and motivational beliefs.

A. Students' motivational beliefs in physics and other STEM fields

The expectancy-value theory (EVT) [49,50] is one of the most prominent approaches to the study of students' motivational beliefs. In the EVT, expectancy refers to students' belief in their ability to succeed in a given task [50]. Value refers to the subjective task value for students, which can be differentiated into four components: intrinsic value, attainment value, utility value, and cost [50]. Intrinsic value refers to students' interest in the task and the enjoyment they experience from performing the task. Attainment value reflects how important students themselves feel it is for them to develop mastery and do a good job in the field [50]. Utility value pertains to students' perception of whether the task can help them achieve some other goals [50]. The last value component is cost, which refers to the assessments of how much effort and time will be taken to engage in the task as well as the amount of opportunity cost and stress caused by the task [50]. In the EVT, students' learning goals, academic engagement and performance, and persistence in a field are impacted by their expectancy of success and the four components of value [50].

The expectancy component of EVT is closely related to the concept of self-efficacy in Bandura's social cognitive theory, which is defined as one's belief in one's ability to succeed in a specific area or accomplish a task [51,52]. Prior research suggests that self-efficacy is an important motivational belief for students to excel in a domain [4,5,10,13]. Studies have shown that students' engagement and performance can be influenced by their self-efficacy [29,31,35,53,54]. For example, students who have high self-efficacy tend to see difficulties as challenges and believe that productive struggles can help them improve, so they often choose to take challenging courses and ask to do more challenging problems than students with low self-efficacy, who usually see difficulties as threats and obstacles to success [30].

The intrinsic value in EVT is closely related the concept of interest, which refers to students' curiosity, enjoyment and engagement in a specific area [55,56]. Studies have shown that interest can also influence students' learning [28,29,56–60]. For example, one study showed that students' performance can be improved by connecting physics courses to students' daily lives or using evidence-based curricula to make the courses more engaging and interesting [61].

In addition, students' identity in a specific field such as physics is another important motivational belief that influences their career decisions [62–69]. Students' physics identity is related to whether they see themselves as a physics person [1–3,62,65]. Some studies have found that female students often report lower physics identity than male students [2,70,71]. This gender difference in physics identity has been shown to be related to societal biases and stereotypes about who belongs in and can succeed in physics [72–74]. These stereotypes can negatively influence women's experiences, which may lower their identity and lead to withdrawal from physics [36,75,76]. Therefore, investigating students' physics identity may help us understand the gender difference in participation in physics. We now turn to our theoretical framework for investigating the effect of students' perception of the inclusiveness of the learning environment on their motivational beliefs and academic performance.

II. THEORETICAL FRAMEWORK

A. Disciplinary identity theories

In Carlone and Johnson's science identity framework [1], students' science identity includes three interrelated constructs: competence (belief in one's competence), performance (belief in ability to perform), and recognition (recognition of self and by others as a "science person"). Hazari *et al.* adapted this model to physics and added interest to this model [3]. In addition, Hazari *et al.* developed quantitative measures for these constructs and found that competence and performance factored into a single construct [3]. Moreover, they separated recognition of self and by others and used a single item ("I see myself as a physics person") to measure students' overall physics identity [3]. Potvin and Hazari noted that "this single item ("I see myself as a physics person") is not intended to measure the totality of the nuance and meaning of students' physics identity; rather in this case we have found, as previously, that this item acts as an excellent and simple stand-in for students' self-perceptions about physics" [77]. This is consistent with prior studies suggesting that a single-item indicator is reasonable when representing global constructs or when a holistic impression is desired [78,79]. In Hazari *et al.*'s later studies using structural equation modeling, they found that students' overall physics identity was predicted by their interest, competence or performance beliefs, and perceived recognition from other people [65,80–82]. This physics identity framework has been used to study physics identity of students in high school physics classes [83,84] as well as college students with a variety of majors [70,81,85–87].

The definition of physics competence or performance beliefs is peoples' beliefs about their ability to understand and perform physics [3], which is very similar to the definition of self-efficacy for the purposes of our research, which uses validated survey data, and our survey items were adapted from prior studies that use the term self-efficacy [88,89]. Moreover, prior studies have shown that self-efficacy is also an important predictor of students' overall identity [90,91]. Therefore, in this study, we will use the physics identity model in which overall physics identity is predicted by self-efficacy, interest, and perceived recognition.

B. Factors that can affect self-efficacy and interest

According to Bandura's social cognitive theory [92,93], there are four factors that can affect people's self-efficacy: mastery experiences (personal experiences of success or failure), vicarious experiences (observing other people's experiences of success or failure), social persuasion (encouragement or discouragement from other people), and physiological states (e.g., anxiety and depression can decrease self-efficacy). Therefore, interaction with other people and recognition from other people play a very important role in students' self-efficacy development. In addition, prior studies have shown that people's physiological states are closely related to their sense of belonging in an environment [94,95] For example, lack of sense of belonging has been shown to contribute anxiety and depression [95].

According to the four-phase model of interest development developed by Hidi and Renninger [96], there are four phases in the development of individual's interest in a certain object. The first stage is triggered situational interest. In this stage, students' interest can be triggered by environmental features such as novelty and surprise. The second stage is maintained situational interest. In this stage, situational interest is held and sustained through meaningfulness of tasks and/or personal involvement. The third stage is emerging individual interest. In this stage, students value the opportunity to reengage tasks related to their emerging triggered situational interest and will opt to do these if given a choice. The last stage is well-developed individual interest. In this stage, students have relatively enduring predisposition to reengage with particular classes of content overtime. Prior studies have shown that students' situational interest is mainly triggered and maintained by external factors [97,98]. For examples, learning environments that provide meaningful and personally engaging activities, such as cooperative group work and one-on-one tutoring can contribute to the maintenance of situational interest [99,100]. In the third and fourth stages, even though students have predisposition to reengage with the content, prior studies have shown that learners with individual interest also need encouragement from others to persevere when confronted with difficulty [101,102]. As we can see, in each stage, learning environment and interaction with other people are very important for students to develop and sustain interest.

C. Factors that may contribute to the gender difference on concept inventories in physics

Prior studies have shown that in addition to motivational beliefs, students' academic performance can also be influenced by their learning environment [103,104]. We note that in physics, students' physics conceptual understanding is an important academic outcome. However, prior studies showed that female students often have lower average scores than male students on physics concept inventories [105–109]. For example, a prior study showed that men, on average, outperform women on the mechanics conceptual inventories by 13% on the pretest and by 12% on the posttest [17]. Many studies exploring gender differences in physics conceptual performance has been conducted with the Force Concept Inventory (FCI), which is one of the most commonly used concept inventories in physics for introductory mechanics [110]. 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 [111]. In addition, some other studies show that particular items on the FCI may be biased against women or men [112,113]. Other factors such as students' academic achievement [114], scientific reasoning ability [115,116], and psychological factors [117,118] 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 [7,119,120]; however, this effect has not been consistently reproduced in other studies [12,20,121]. In particular, a study shows that in a noninclusive learning environment, female students may benefit less from interactive learning because they do not feel safe to express themselves, and thus the gender gap may be even larger than in a traditional lecture-based course [122].

D. Students' perception of the inclusiveness of the learning environment

As discussed above, learning environment plays an important role in developing students' motivational beliefs and improving their physics conceptual understanding. However, to our knowledge, no prior studies have quantitively investigated the effect of students' perception of the inclusiveness of the learning environment on their physics conceptual understanding. Therefore, in this study, we focus on how students' perception of the inclusiveness of learning environment predicts their physics conceptual understanding measured by FCI and their motivational beliefs in a college level calculus-based introductory mechanics course. Similar to many quantitative studies in physics education research designed to examine the relations between students' attributes and learning environments [123], there are several ontological assumptions behind our study. The first ontological assumption is that students' attributes such as their motivational beliefs and physics conceptual understanding are a function of different features of the learning environment in which students are placed. Thereby, a change in the environment, if systematically varied, may lead to a change in the students' attributes [123]. Second ontological assumption is that students' perception of the inclusiveness of the learning environments is composed of different components, whose effects on students' attributes can be investigated and quantified using appropriate research methods. Students' perception of the inclusiveness of the learning environment is based on their interactions with people in the learning environment such as their instructors or TAs and peers. For example, whether students feel validated and recognized by other people and whether the interactions with others are meaningful and enjoyable. As discussed earlier, these factors are important for students to develop interest and self-efficacy [52,56]. Similarly, prior studies showed that students' perceived recognition from other people also predicts their overall physics identity and academic performance [71,124]. Therefore, in this study, we include students' perceived recognition from others and their perception of the effectiveness of the peer interaction as two components of students' perception of the inclusiveness of the learning environment.

Another important component of students' perception of the inclusiveness of the learning environment is sense of belonging, which is the feeling of inclusion or acceptance into a group of people [7,75,76,125–128]. Compared with perceived recognition and perception of the effectiveness of the peer interaction, students' sense of belonging directly reflects their overall feeling of the inclusiveness of the learning environment. Prior studies have shown that students' sense of belonging is closely related to their motivational beliefs and academic performance. For instance, Freeman and colleagues [129] found that students' sense of belonging in a specific college class was positively associated with their self-efficacy, task utility, and intrinsic motivation. In physics, prior studies have shown that students' sense of belonging predicts their overall physics identity, perceived utility value of academic tasks, and course grades [82,130,131].

Therefore, in this study, we include sense of belonging, perceived recognition, perception of the effectiveness of the peer interaction as three components of students' perception of the inclusiveness of the learning environment and investigate how they predict students' physics conceptual understanding and motivational beliefs at the end of a calculus-based introductory physics course. Another novelty of our study is that we focus on the net effect of each component of the inclusiveness of learning environment on the course outcomes by controlling for the effects of the other two (which will be discussed in detail in the next section). This is important because these three inclusiveness of learning environment constructs have been shown to correlate with each other [82,125,132], and by controlling for the effects of potential confounding variables, the net effect can tell us how much effect a predictor has on an outcome construct beyond the effects of other predictors.

III. THE PRESENT STUDY AND ANALYTICAL FRAMEWORK

Inspired by the above studies, we conducted a study focusing on students' physics motivational beliefs and conceptual understanding in a calculus-based introductory mechanics course at a large public university. We investigated how students' perception of the inclusiveness of the learning environment (including students' sense of belonging, perceived effectiveness of peer interaction, and perceived recognition) predicts their motivational beliefs and FCI scores at the end of the course after controlling for students' motivational beliefs and FCI scores at the beginning of the course as well as their high school GPA and scholastic assessment test (SAT) math scores. The SAT is a standardized test widely used for college admissions in the United States, which includes math and verbal sections. For convenience, perceived effectiveness of peer interaction is shortened to peer interaction in the rest of the paper. We note that the learning environment here is not only the classroom environment but also includes students' experiences outside the class. For example, students may work together on their homework after class, and they could also ask for help during TAs' or instructors' office hours or communicate with the instructor or TA via email about various issues pertaining to the course. As shown in Fig. 1, the thirteen constructs are divided into three groups: what we control for, students' perception of the inclusiveness of the learning environment, and outcomes. Students' gender, SAT math, high school GPA (HS GPA), and their self-efficacy, interest, and FCI scores at the beginning of the course (Pre SE, Pre Interest, and Pre FCI) are constructs that we control for. Outcomes include students' self-efficacy, interest, FCI scores and overall physics identity at the end of the course (Post SE, Post Interest, and Post FCI). Perceived recognition (Perceived Recog), peer interaction (Peer Int) and sense of belonging (Belonging) constitute the perception of the inclusiveness of learning environment.

In our study, students' peer interaction, perceived recognition, sense of belonging and overall physics identity were measured at the end of the course because only after the course can students answer these survey questions based on their real experience in the course such as their interaction with peers, TAs and instructors. It is expected that students' responses to the survey in pre- and postsurvey are correlated because they are students' responses to the same questions pertaining to the same construct at two different time points. However, if students' motivational

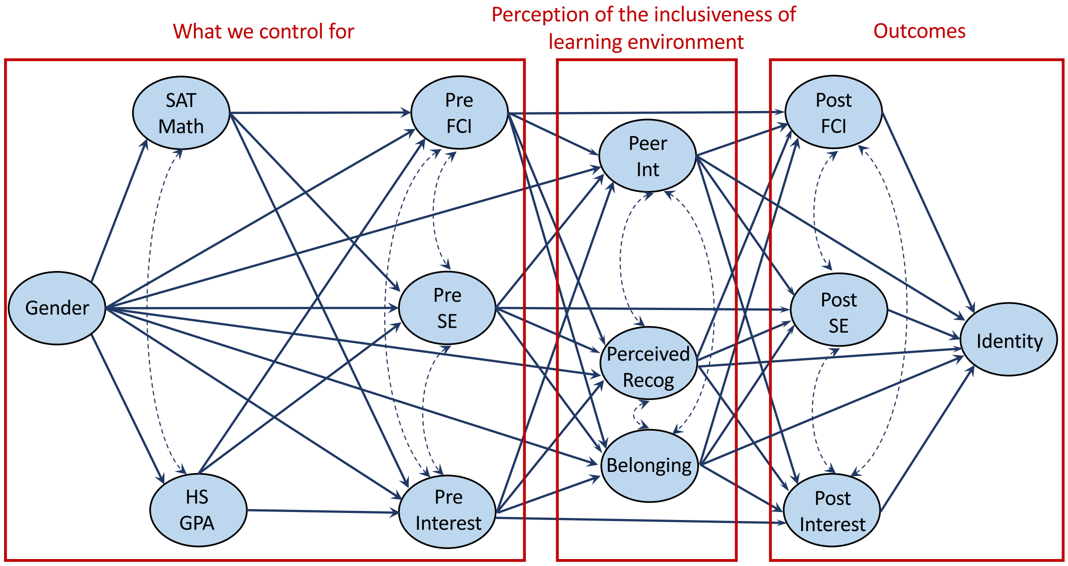

FIG. 1. Schematic representation of the theoretical model in which the relation between gender and overall physics identity is mediated through SAT Math scores, high school GPA (HS GPA), and FCI scores as well as peer interaction (Peer Int), perceived recognition (Recog), sense of belonging, self-efficacy (SE), and interest. The solid lines represent regression paths, and the dashed lines represent covariances. From left to right, all possible regression paths were considered, but only some of the paths are shown here for clarity.

beliefs changed from pre to post, we want to study whether the inclusiveness of the learning environment helps to explain the changes and what role is played by each construct in the inclusiveness of learning environment.

In this study, we first investigated how students' self-efficacy, interest, and FCI scores changed from the beginning to the end of the course and whether there were gender differences in the constructs studied. Then, we used structural equation modeling (SEM) to study how students' perception of the inclusiveness of learning environment predicts students' self-efficacy, interest, overall physics identity and FCI scores at the end of the course. To better understand the role played by each inclusiveness of learning environment construct, we first considered a model with perceived recognition as the only inclusiveness of learning environment construct to analyze how much variance in the outcome constructs is explained by the model. Then, we added peer interaction and sense of belonging into this model one by one to investigate whether adding these constructs helps to explain extra variance in the outcome constructs.

We note that our research design is guided by several epistemological commitments [123]. First, in this study, we made the decision to focus on students' perception of the inclusiveness of the learning environment, which could be different from the perceptions of instructors/TAs or a third party who observes the course. However, since students are the ones who go through the learning experiences, we believe it is important to study students' point of view about the inclusiveness of the learning environment. Second, the three components (perceived recognition, sense of belonging, and perception of peer interaction) cover different aspects of students' perception of the inclusiveness of the learning environment with regard to interactions with others and an overall belonging we want to investigate. Other factors such as level of anxiety could also contribute to students' perception of the inclusiveness of the learning environment; however, since other factors often strongly correlate with the three components already included, we did not include them in our model. Third, by using statistical methods such as SEM with large sample size, our aim is to investigate the relationships between the constructs studied. Because of the nature of quantitative studies, our study will show the trends and patterns in our data rather than focusing on any individual student.

IV. RESEARCH QUESTIONS

In this study, we used quantitative methods to investigate how students' perception of the inclusiveness of the learning environment predicts physics motivational beliefs and performance on the Force Concept Inventory in a calculus-based introductory physics sequence at a large state-related university in the U.S. This course is mandatory for students majoring in engineering, physical science, and mathematics in their first year at the university. Specifically, we address the following research questions:

RQ1. Are there gender differences in students' FCI scores and motivational beliefs and do they change from pre to post?

RQ2. How do the components of students' perception of the inclusiveness of learning environment (including sense of belong, peer interaction and perceived recognition) predict students' self-efficacy, interest, overall physics identity, and FCI scores at the end of the course after controlling for students' gender, high school GPA, SAT math scores, and their self-efficacy, interest and FCI scores at the beginning of the course?

RQ3. Does gender moderate the relationship between any pairs of constructs in the models (i.e., does the strength of relationship given by the standardized regression coefficients between any two constructs in the models differ for women and men)?

RQ4. If gender does not moderate any path in the model, how does gender predict

a. the factors that were controlled for?
b. the inclusiveness of learning environment constructs after controlling for students' high school GPA, SAT math scores, and their self-efficacy, interest and FCI scores at the beginning of the course?
c. the learning outcomes after controlling for everything in the model?

RQ5. What role is played by each of the three components we have included in the inclusiveness of learning environment in predicting the outcome constructs?

RQ6. Based on the aspects of students' perception of the inclusiveness of the learning environment that explain most of the variance in the outcome constructs, which model is most productive for providing guidelines for creating an inclusive environment?

V. METHODOLOGY

A. Participants and data sources

The data used in this study were collected from a college level calculus-based introductory physics course in two consecutive school years at a large public research university in the U.S. This course is generally mandatory and taken by engineering, physical science and mathematics majors in the first semester of their first year of undergraduate studies. 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. In addition, the course assessment in this course is largely based on students' performance on the midterm and final exams, which mainly focus on quantitative problem solving. Moreover, there was very little focus on using evidence-based pedagogies or intentional efforts to promote equity and inclusion in this course.

Students' motivational beliefs and perception of the inclusiveness of the learning environment were measured using a validated survey. The Force Concept Inventory [110] was used to measure students' conceptual understanding of introductory mechanics. Both the survey and conceptual test were administered to students in the first and last recitation class of the semester. The demographic data of students—such as gender—were provided by the university. Students' SAT math scores and high school GPA were also obtained from the university records. 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. There were 1364 students participating our study at the beginning of the course and 1203 students at the end of the course. In this study, we focused on 1045 students (382 female students and 663 male students) who completed the survey and FCI test at both the beginning and end of the course (matched students from pre to post) because we want to investigate how students' motivational beliefs and FCI scores change from the beginning to the end of the course and what role is played by students' perception of the inclusiveness of the learning environment in these changes. Some possible reasons that some students did not take the pre- or postsurvey or test include 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-drop period is the first few weeks of the course). There were no missing data in our study except a couple of students forgetting to respond to one survey item. We recognize that gender identity is not a binary construct. However, students' gender information was collected by the university, which offered binary options. For our analysis, we use the binary gender data. Fewer than 1% of the participants did not provide this information and therefore were not included in this analysis.

B. Survey instruments

In this study, our analysis includes three motivational constructs (physics self-efficacy, physics interest, and overall physics identity) and three perceptions of the inclusiveness of the learning environment constructs (peer interaction, perceived recognition, and sense of belonging). The questions for each construct are listed in Table I. The survey questions were adapted from existing motivational research [89,133–138] and were revalidated in our prior work [10,139–142]. The validation and refinement of the survey involved use of one-on-one student interviews with both introductory and advanced students [10,142–144], exploratory and confirmatory factor analysis (EFA and CFA) [145], Pearson correlation between different constructs, and Cronbach's alpha [146,147].

Physics self-efficacy represents students' belief about whether they can perform well in physics. In our survey, we had four items for self-efficacy (Cronbach's alpha = 0.69 for pre-self-efficacy and Cronbach's alpha = 0.80 for post-self-efficacy [147]). These items had the response scale "NO!, no, yes, YES!", which is a 4-point Likert scale (1–4). We also had four items for physics interest

TABLE I. Survey items for each of the motivational constructs. The Cronbach alphas and CFA item loadings (Lambda and p values of the significance test for each item loading) shown here were calculated with postdata.

Construct and item	Lambda	p value
Overall physics identity		
I see myself as a physics person	1.000	<0.001
Physics self-efficacy (Cronbach's alpha = 0.80)		
I am able to help my classmates with physics in the laboratory or in recitation	0.802	<0.001
I understand concepts I have studied in physics	0.828	<0.001
If I study, I will do well on a physics test	0.791	<0.001
If I encounter a setback in a physics exam, I can overcome it	0.723	<0.001
Physics interest (Cronbach's alpha = 0.82)		
I wonder about how physics works[a]	0.685	<0.001
In general, I find physics[b]	0.899	<0.001
I want to know everything I can about physics	0.863	<0.001
I am curious about recent physics discoveries	0.743	<0.001
Physics perceived recognition (Cronbach's alpha = 0.86)		
My family sees me as a physics person	0.920	<0.001
My friends see me as a physics person	0.927	<0.001
My physics TA and/or instructor see me as a physics person	0808	<0.001
Physics sense of belonging (Cronbach's alpha = 0.86)		
I feel like I belong in this class	0.868	<0.001
I feel like an outsider in this class	0.767	<0.001
I feel comfortable in this class	0.877	<0.001
I feel like I can be myself in this class	0.641	<0.001
Sometimes I worry that I do not belong in this physics class	0.780	<0.001
Physics peer interaction (Cronbach's alpha = 0.91)		
My experience and interaction with other students in this class…		
Made me feel more relaxed about learning physics	0.750	<0.001
Increased my confidence in my ability to do physics	0.962	<0.001
Increased my confidence that I can succeed in physics	0.981	<0.001
Increased my confidence in my ability to handle difficult physics problems	0.904	< 0.001

[a]The response options for this question are "never, once a month, once a week, every day."
[b]The response options for this question are "very boring, boring, interesting, very interesting".

(Cronbach's alpha = 0.75 for pre-interest, Cronbach's alpha = 0.82 for postinterest). The question "I wonder about how physics works" had temporal response options: "never, once a month, once a week, every day." whereas the question "In general, I find physics:" had response options "very boring, boring, interesting, very interesting." The remaining two items were answered on the "NO!, no, yes, YES!" scale. By choosing the four options, students will get a score from 1 to 4 accordingly. For example, if a student finds physics very boring, he or she will get one point for this item. The more interest a student has in physics, the higher score the student will get for this item. There is one item for overall physics identity in this survey (I see myself a physics person). This item involved a four-point Likert response on the scale: "strongly disagree, disagree, agree, and strongly agree" and they correspond to 1 to 4 points, respectively [148]. We note that the use of the single item (I see myself as a physics person) to measure students' overall physics identity was adapted from previous studies of Hazari *et al.* [3,80,149]. As noted earlier, prior studies suggest that a single-item indicator is appropriate when representing global constructs or when a holistic impression is desired [78,79]. The single overall identity item (to which degree individuals perceive themselves as a "type of person") has also been commonly adapted to study students' disciplinary identity in many other areas such as math [86], biology [150], chemistry [151], engineering [65], science [152], and STEM overall [153]. Many of these studies showed that the single item for disciplinary identity is highly correlated with students' career pursuits in the corresponding area [77,154], which is aligned with prior research on the relationship between identity and career pursuits. In addition, prior studies showed that the single overall identity item is also highly correlated with the weighted composite score of the components of identity (such as self-efficacy, interest, and perceived recognition) [150,153,154]. Therefore, in this study, we used this single item as holistic measure of students' physics identity.

In addition, perceived recognition, peer interaction and sense of belonging are the perception of the inclusiveness of the learning environment constructs in our study. Unlike self-efficacy, interest and overall identity, these three constructs are directly related to students' interactions and experience in the course. Perceived recognition included three items which represent whether a student thinks other people see them as a physics person [2,3,155] (Cronbach alpha's = 0.86). Peer interaction, including four items, represents whether students have a productive and enjoyable experience when working with peers (Cronbach's alpha = 0.91). Both perceived recognition and peer interaction have a four-point Likert response on the scale: "strongly disagree, disagree, agree, and strongly agree." Sense of belonging is about students' overall feelings of whether they belonged in the physics class [127], and it included five items that were scored on a 5-point Likert scale: "not at all true, a little true, somewhat true, mostly true and completely true" (Cronbach alpha = 0.86). Two sense of belonging items ("I feel like an outsider in this class" and "Sometimes I worry that I do not belong in this physics class") were reverse coded, which means that a higher score in these two items represents a lower sense of belonging. A student's score for each construct is the average score of all items in this construct.

C. Quantitative analysis

In this study, we first calculated the mean score for each construct for each student. We note that in our previous study [124], we checked the response option distances for our survey constructs by using item response theory (IRT) to support the use of means across ratings [156,157]. Even for this study, we performed IRT with the new data set 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 [158,159]. Some items have response scales of "strongly disagree, disagree, agree, and strongly agree," some items have response scale "NO!, no, yes, YES!", and the other items have 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. For the first group—strongly disagree, disagree, agree, and strongly agree—the difference between the location parameters were 1.23 and 1.56. For the second group—"NO!, no, yes, YES!"—the difference between the location parameters were 1.32 and 1.98. For the third group—"not at all true, a little true, somewhat true, mostly true, and completely true," the difference between the location parameters were 0.88 and 0.94. These results show that the numerical values for the location differences for item responses are comparable, which suggests that calculating the traditional mean score of items is reasonable [156,159]. Furthermore, we estimated the IRT-based scores with expected *a posteriori* (EAP) computation method for each construct. The results show that the correlation coefficients between the mean scores and the IRT-based scores are >0.95 for all constructs studied, which also indicates that the use of mean scores is reasonable [156].

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 (see Appendices A and B), which is important for choosing appropriate analysis method [160,161]. The distributions of academic data (including high school GPA, SAT math score, and pre- and post-FCI score) are presented via graphs (Appendix A), and the distributions of students' responses to the Likert scale survey items are presented via tables (Appendix B). 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 normality assumption is not satisfied or the data are ordinal [162]. We used the Wilcoxon signed-rank test to estimate the changes in students' responses to the survey 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 normality assumption is not satisfied or the data are ordinal [162].

Then, we used the R [163] software package "lavaan" to conduct structural equation modeling [164] to study how students' perception of the inclusiveness of learning environment predicted their motivational beliefs and FCI scores at the end of the course after controlling for students' gender, high school GPA and SAT math as well as their motivational beliefs and FCI scores at the beginning of the course. SEM is a multivariate statistical analysis technique that is used to model the relations between measured 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 measured 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. 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, which allowed us to calculate the total regression effect by adding the direct and indirect regression coefficients up.

The assumptions associated with SEM include: correct model specification, sufficiently large sample size, and no systematic missing data [165–167]. Our study is based on the identity model in which students' overall physics identity is predicted by their perceived recognition, self-efficacy, and interest. This model has been examined by many prior studies [3,65,82,124]. According to Kline, a typical sample size in studies where SEM is used is about 200 [165], so the sample size of our study ($N = 1045$) is sufficiently large for SEM. Moreover, since we focus on students who participated both presurvey and postsurvey (matched students from pre to post), there were no missing data in our study except a couple of students forgetting to respond to one survey item. In addition, a well fitted measurement model (which is also called confirmatory factor analysis) is also very important for performing full SEM [168]. As we will discuss in the next paragraph, our data fit the measurement model very well. Moreover, Table I shows that almost all factor loadings are higher than 0.7, which is considered as satisfactory [168]. This means that the constructs extract sufficient variance from the observed variables, which allows us to perform full SEM [169]. In this study, we used diagonally weighted least square (DWLS) to estimate parameters. DWLS estimation is commonly used to analyze ordinal variables and has also been shown to produce unbiased parameters estimates with great statistical power for non-normal data [170,171].

As noted earlier, the SEM includes two parts: confirmatory factor analysis and path analysis. First, we performed the CFA for each construct. When performing the CFA model, the program automatically fixed the unstandardized loading of the first specified indicator for each construct to 1.0 to assign the corresponding factor a scale. Similarly, the unstandardized loading of the single indicator for overall physics identity was also automatically fixed to 1.0, which is consistent with the suggestions of Kline [165]. As Kline noted, by fixing the unstandardized loading to 1.0, the scale of the latent variable is set equal to the scale of the indicator variable. The model fit is good if the fit parameters are above certain thresholds. In CFA, comparative fit index $(\mathrm{CFI}) > 0.9$, Tucker-Lewis index $(\mathrm{TLI}) > 0.9$, root mean square error of approximation $(\mathrm{RMSEA}) < 0.08$ and standardized root mean square residual $(\mathrm{SRMR}) < 0.08$ are considered acceptable and $\mathrm{RMSEA} < 0.06$ and $\mathrm{SRMR} < 0.06$ are considered a good fit [172]. In our study, $\mathrm{CFI} = 0.979$, $\mathrm{TLI} = 0.975$, $\mathrm{RMSEA} = 0.051$, and $\mathrm{SRMR} = 0.044$, which represents a good fit. This result provides quantitative support for us to organize the motivational constructs as proposed.

Before performing the path analysis, we calculated the pairwise correlations between each pair of constructs (see Table II) [146]. 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 [173,174]. As shown in Table II, there are relatively strong correlations among students' motivational

TABLE II. Correlations between the constructs in the mediation model.

Observed variable	1	2	3	4	5	6	7	8	9	10	11	12
1. SAT math	…	…	…	…	…	…	…	…	…	…	…	…
2. HS GPA	0.22	…	…	…	…	…	…	…	…	…	…	…
3. Overall physics identity	0.10^{**}	-0.08^{*}	…	…	…	…	…	…	…	…	…	…
4. Pre-FCI	0.36	0.11	0.42	…	…	…	…	…	…	…	…	…
5. Pre-self-efficacy	0.16	0.07^{*}	0.57	0.36	…	…	…	…	…	…	…	…
6. Pre-interest	-0.01^{ns}	-0.11^{**}	0.64	0.31	0.62	…	…	…	…	…	…	…
7. Post-FCI	0.33	0.12	0.38	0.77	0.34	0.27	…	…	…	…	…	…
8. Post-self-efficacy	0.19	0.03^{ns}	0.72	0.49	0.61	0.44	0.45	…	…	…	…	…
9. Post-interest	-0.01^{ns}	-0.15	0.75	0.32	0.48	0.90	0.31	0.62	…	…	…	…
10. Perceived recognition	0.15	-0.01^{ns}	0.88	0.44	0.52	0.58	0.41	0.74	0.68	…	…	…
11. Peer interaction	0.14	0.03^{ns}	0.54	0.25	0.40	0.31	0.23	0.68	0.46	0.52	…	…
12. Sense of belonging	0.22	0.05^{ns}	0.65	0.40	0.46	0.38	0.39	0.80	0.53	0.65	0.68	…

p values are indicated by ** for $0.001 \le p < 0.01$, * for $0.01 \le p < 0.05$, and $^{\mathrm{ns}}$ for $p > 0.05$. All the other correlation coefficients have $p < 0.001$.

beliefs, while the correlation between motivational beliefs and SAT math or high school GPA are relatively small. We note that in Table II, there are several very strong correlations. For example, the correlation coefficient between overall physics identity and perceived recognition is 0.88, which is consistent with Godwin *et al.* and Kalender *et al.*'s prior work [65,124] showing that perceived recognition is the largest predictor of overall physics identity. Another large correlation coefficient is between students' post-self-efficacy and sense of belonging, which is 0.80. According to prior work done by Kalender *et al.*, these two constructs are indeed strongly correlated with each other even though they are separate constructs [175].

To analyze the relations among the constructs, we performed the path analysis. The path analysis in SEM gives regression coefficients β for paths between each pair of constructs and the value of each β is a measure of the strength of that relationship. We first analyzed the saturated SEM model that includes all of the possible links between different constructs, and then we used the modification indices to improve the model fit. We kept path links which were statistically significant in SEM path analysis. Before performing gender mediation analysis, we first tested the gender moderation relations between each pair of constructs using multigroup SEM (to investigate any interaction effects with gender), which includes testing of factor loadings, indicator intercepts, residual variances, and regression coefficients. Results showed that in all our models, strong measurement invariance holds and there is no difference in any regression coefficients by gender, which allowed us to perform the gender mediation analysis using SEM (see Appendix C for detailed multigroup SEM analysis results).

Because many quantitative studies have shown that perceived recognition is a strong predictor of students' motivational beliefs and overall physics identity [124,176–178], all of the models shown in this paper include perceived recognition as one of the inclusiveness of the learning environment constructs. To understand the role played by each inclusiveness of learning environment construct, we first considered a model (model 1) with perceived recognition as the only inclusiveness of learning environment construct to investigate how students' motivational outcomes and FCI scores at the end of the course are predicted by it. Then, in model 2 we added peer interaction and in model 3 we added sense of belonging as additional constructs in the inclusiveness of learning environment to study whether adding these constructs helps to explain extra variance in the outcome constructs compared with model 1. Finally, we included all three inclusiveness of learning environment components in our model (model 4) to study how each component predicts the course outcomes after controlling for the effects of the other two components. Moreover, we compared the model fit indices of the four models and the variance in each outcome construct explained by each model to understand the role played by each inclusiveness of learning environment component and to determine if all three components are productive.

VI. RESULTS

A. Gender differences in students' motivational beliefs and FCI scores

Table III shows the descriptive statistics of students' physics interest, physics self-efficacy, 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 for Wilcoxon rank-sum tests and Wilcoxon signed-rank tests [179]. As shown in Table III, female students had significantly lower average interest, self-efficacy, and FCI scores than male students, and the effect size of gender difference in self-efficacy increased from 0.15 to 0.24 by the end of the course. In addition, Table III shows that both male and female students' interest and self-efficacy dropped generally from pre to post, and the decrease in female students' interest and self-efficacy dropped (effect size is −0.21 for interest −0.29 for self-efficacy) even more than male students' (effect size is −0.16 for interest and −0.17 for self-efficacy). Even though both female and male students' FCI scores increased by the end of the course, the gender difference is maintained.

Table IV shows the descriptive statistics of students' perception of the inclusiveness of the learning environment (including peer interaction, perceived recognition, and sense of belonging) and overall physics identity. As shown in Table IV, female students had significantly lower average scores in all of the four constructs than male students. These results indicate that, in the current learning environment, female students reported less benefit from peer interaction and also felt a lower sense of belonging than male students. Moreover, female students' average scores pertaining to perceived recognition and overall physics identity indicate that on average, female students did not think others see them as a physics person, and they did not see themselves as a physics person either. In Appendix B, we report the percentages of students who selected each choice for each survey item, which show consistent results with the descriptive statistics shown in Tables III and IV.

Table V shows the descriptive statistics of students' high school GPA and SAT math scores. As shown in Table V, there was no statistically significant gender difference in students' SAT math scores, and female students had a higher average high school GPA than male students.

B. SEM path models

In this section, we describe results of the structural equation modeling carried out to investigate how students'

TABLE III. Descriptive statistics of pre- and postinterest, self-efficacy (SE), and FCI scores for female and male students, 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 for Wilcoxon rank-sum tests and Wilcoxon signed-rank tests [179]. Hake suggested that values of $g < 0.3$, $0.3 < g < 0.7$, and $g > 0.7$ represent small, medium, and large normalized gains [180]. A minus sign indicates that students' average score decreased from pre to post.

	Pre-Interest (1–4)	Post-Interest (1–4)	Statistics		Pre-SE (1–4)	Post-SE (1–4)	Statistics	
Gender	Mean	Mean	Effect size	p value	Mean	Mean	Effect size	p value
Male	3.19	3.07	−0.16	<0.001	3.12	2.98	−0.17	<0.001
Female	2.89	2.73	−0.21	<0.001	2.96	2.70	−0.29	<0.001
p value	<0.001	<0.001			<0.001	<0.001		
Effect size	0.25	0.26			0.15	0.24		

	Pre-FCI	Post-FCI	Statistics		
Gender	Mean	Mean	Normalized gain (g)	Effect size	p value
Male	62%	73%	0.29	0.45	<0.001
Female	47%	60%	0.25	0.48	<0.001
p value	<0.001	<0.001			
Effect size	0.33	0.30			

TABLE IV. Descriptive statistics of peer interaction, perceived recognition, sense of belonging, and overall physics identity for female and male students, along with the results of Wilcoxon rank-sum tests for gender differences.

Gender	Peer interaction (1–4)	Perceived recognition (1–4)	Sense of belonging (1–5)	Overall physics identity (1–4)
Male	2.97	2.58	3.73	2.62
Female	2.70	2.26	3.36	2.19
p value	<0.001	<0.001	<0.001	<0.001
Effect size	0.20	0.21	0.20	0.24

perception of the inclusiveness of the learning environment predicts their motivational beliefs and FCI scores at the end of the course. As noted earlier, we first considered a model (model 1) in which perceived recognition was the only inclusiveness of learning environment construct. Then we added peer interaction (model 2) or sense of belonging (model 2) to the inclusiveness of learning environment one by one to analyze how each helped to predict students' self-efficacy, interest, overall physics identity, and FCI scores at the end of the course. Finally, we included all three constructs in our model (model 4) and studied how these constructs mediated the outcomes together and what role was played by each of them.

TABLE V. Descriptive statistics of female and male students' high school GPA and SAT math scores, along with the results of Wilcoxon rank-sum tests for gender differences. A minus sign indicates that female students have a higher average score than male students.

	Mean			
Grades (Score Range)	Male	Female	p value	Effect size
High school GPA (0-5)	4.10	4.25	<0.001	−0.19
SAT math (400-800)	701	694	0.188	0.04

1. Model 1: Perceived recognition

In our first model (model 1), perceived recognition is the only inclusiveness of learning environment construct. The path analysis results of the SEM model are presented visually in Fig. 2. The model fit indices suggest a good fit to the data: $\mathrm{CFI} = 0.987$ (>0.90), $\mathrm{TLI} = 0.986$ (>0.90), $\mathrm{RMSEA} = 0.051$ (<0.08) and $\mathrm{SRMR} = 0.053$ (<0.08). The solid lines represent regression paths and the numbers on the lines are regression coefficients (β values), which represent the strength of the regression relations. As shown in Fig. 2, perceived recognition directly predicts students' FCI scores, self-efficacy, interest, and overall physics identity at the end of the course. The direct effect of perceived recognition on post-self-efficacy ($\beta = 0.48$) is even larger than that of pre-self-efficacy ($\beta = 0.27$). In addition, we note that even though pre-self-efficacy directly predicts post-self-efficacy, there is also an indirect path from pre-self-efficacy to post-self-efficacy mediated

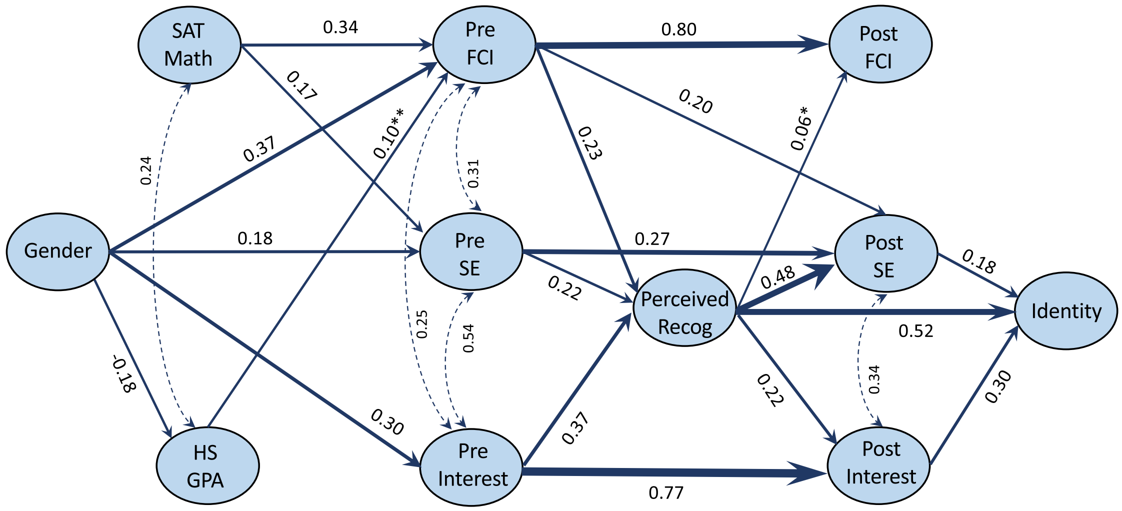

FIG. 2. Schematic diagram of the path analysis part of the structural equation modeling (model 1) between gender and overall physics identity through SAT Math scores, high school GPA (HS GPA), and FCI scores as well as perceived recognition (Recog), self-efficacy (SE), and interest. The solid lines represent regression paths and the dashed lines represent residual covariances. The regression line thickness corresponds to the magnitude of β value (standardized regression coefficient) with $0.01 \leq p < 0.05$ indicated by * and $0.001 \leq p < 0.01$ indicated by **. All the other regression lines show relations with $p < 0.001$.

through perceived recognition. The regression coefficient of the indirect path can be calculated by multiplying the regression coefficients from pre-self-efficacy to perceived recognition ($\beta = 0.22$) and the regression coefficient from perceived recognition to post-self-efficacy ($\beta = 0.48$), which gives us $0.22 \times 0.48 = 0.11$. Similarly, the direct effect from pre-interest to postinterest is $\beta = 0.77$, and the indirect effect is $0.37 \times 0.22 = 0.08$ Consistent with Godwin *et al.* and Kalender *et al.*'s prior work [65,124], Fig. 2 shows that overall physics identity is mainly predicted by self-efficacy, interest, and perceived recognition, and perceived recognition is the largest predictor. In addition, perceived recognition also predicts students' post-FCI scores even after controlling for their pre-FCI scores, high school GPA, and SAT math scores. We note that gender directly predicts high school GPA with a negative regression coefficient ($\beta = -0.18$), which means that female students on average had a somewhat higher high school GPA than male students. This is consistent with the results shown in Table V.

2. Model 2: Perceived recognition and peer interaction

In the second model (model 2), we include both perceived recognition and peer interaction in the perception of the inclusiveness of the learning environment. The results of the SEM model are presented visually in Fig. 3. This model also fits the data very well. CFI = 0.988 (>0.90), TLI = 0.987 (>0.90), RMSEA = 0.047 (<0.08), and SRMR = 0.051 (<0.08). The results show that students' peer interaction directly predicts their post-self-efficacy ($\beta = 0.38$) and postinterest ($\beta = 0.18$), and it also mediates the effect from pre-self-efficacy to post-self-efficacy with indirect regression coefficient $0.37 \times 0.38 = 0.14$. We note that the direct effects of perceived recognition on post-self-efficacy and postinterest are weaker in model 2. This is because the regression coefficient from a predictor to an outcome represents the expected change in the outcome as a result of change in the predictor in standard deviation units while controlling for the correlated effects of other predictors [181]. Since there is a shared variance between peer interaction and perceived recognition, after peer interaction was added to the model, the correlated effect of peer interaction is controlled for when estimating the regression coefficients from perceived recognition to post-self-efficacy and postinterest, so the regression coefficients decreased. We note that the direct effect of perceived recognition on post-FCI becomes statistically insignificant in model 2. On the other hand, the regression coefficients from perceived recognition, post-self-efficacy, and postinterest to overall physics identity are similar to those in model 1.

3. Model 3: Perceived recognition and sense of belonging

We next analyzed a SEM model (model 3) which includes only perceived recognition and sense of belonging as the inclusiveness of learning environment constructs. The results of the SEM model are presented visually in Fig. 4. The model also fits the data well [CFI = 0.979 (>0.90), TLI = 0.977 (>0.90), RMSEA = 0.054 (<0.08), and SRMR = 0.053 (<0.08)]. As shown in Fig. 4, students' sense of belonging directly predicts their post-FCI scores, post-self-efficacy, and postinterest. Similarly, because there is a correlation between sense of belonging

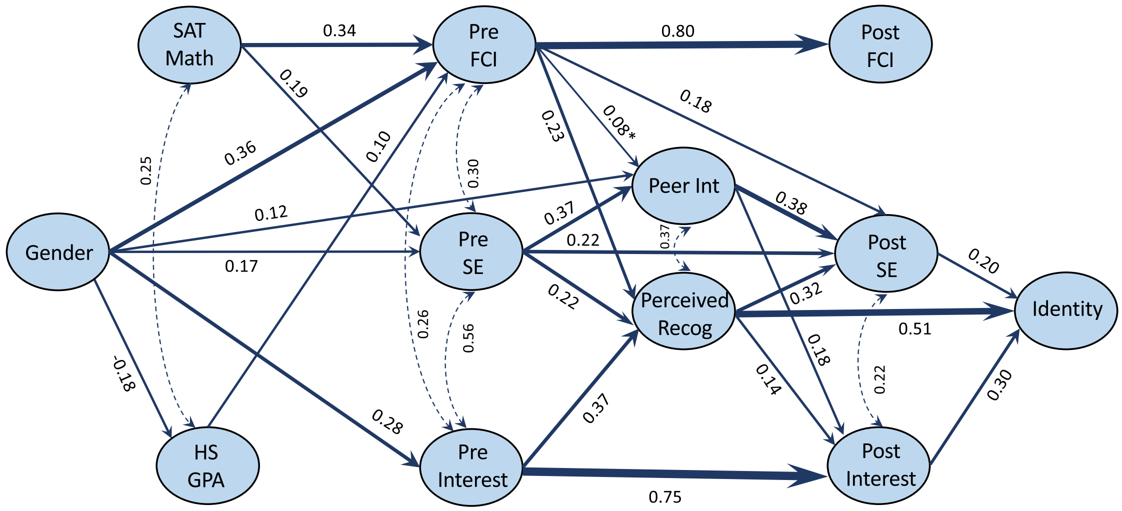

FIG. 3. Schematic diagram of the path analysis part of the structural equation modeling (model 2) between gender and overall physics identity through SAT Math scores, high school GPA (HS GPA), and FCI scores as well as peer interaction (Int), perceived recognition (Recog), self-efficacy (SE), and interest. The solid lines represent regression paths and the dashed lines represent residual covariances. The regression line thickness corresponds to the magnitude of β value (standardized regression coefficient) with $0.01 \leq p < 0.05$ indicated by * and $0.001 \leq p < 0.01$ indicated by **. All the other regression lines show relations with $p < 0.001$.

and perceived recognition, the correlated effect of sense of belonging was controlled for when estimating the regression coefficients from perceived recognition to the outcome constructs, and thus the direct effects of perceived recognition on post-FCI scores, post-self-efficacy, and post-interest became weaker or insignificant compared with those in model 1. On the other hand, the regression coefficients from perceived recognition, post-self-efficacy, and post-interest to overall physics identity are also similar to those in models 1 and 2.

4. Model 4: Perceived recognition, peer interaction, and sense of belonging

Finally, we consider a SEM model (model 4) which includes all three inclusiveness of learning environment constructs. Figure 5 shows the results visually. The model also fits the data very well [CFI = 0.982 (>0.90), TLI = 0.981 (>0.90), RMSEA = 0.049 (<0.08) and SRMR = 0.051 (<0.08)]. As shown in Fig. 5, post-self-efficacy is directly predicted by all three inclusiveness of learning environment constructs, and sense of belonging is the

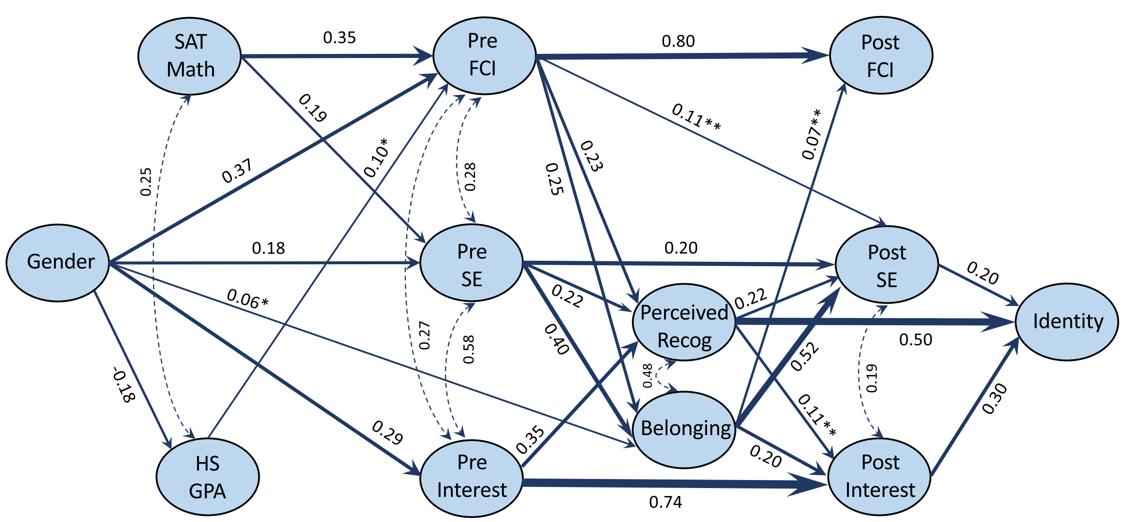

FIG. 4. Schematic diagram of the path analysis part of the structural equation modeling (model 3) between gender and overall physics identity through SAT Math scores, high school GPA (HS GPA), and FCI scores as well as perceived recognition (Recog), sense of belonging, self-efficacy (SE), and interest. The solid lines represent regression paths and the dashed lines represent residual covariances. The regression line thickness corresponds to the magnitude of β value (standardized regression coefficient) with $0.01 \leq p < 0.05$ indicated by * and $0.001 \leq p < 0.01$ indicated by **. All the other regression lines show relations with $p < 0.001$.

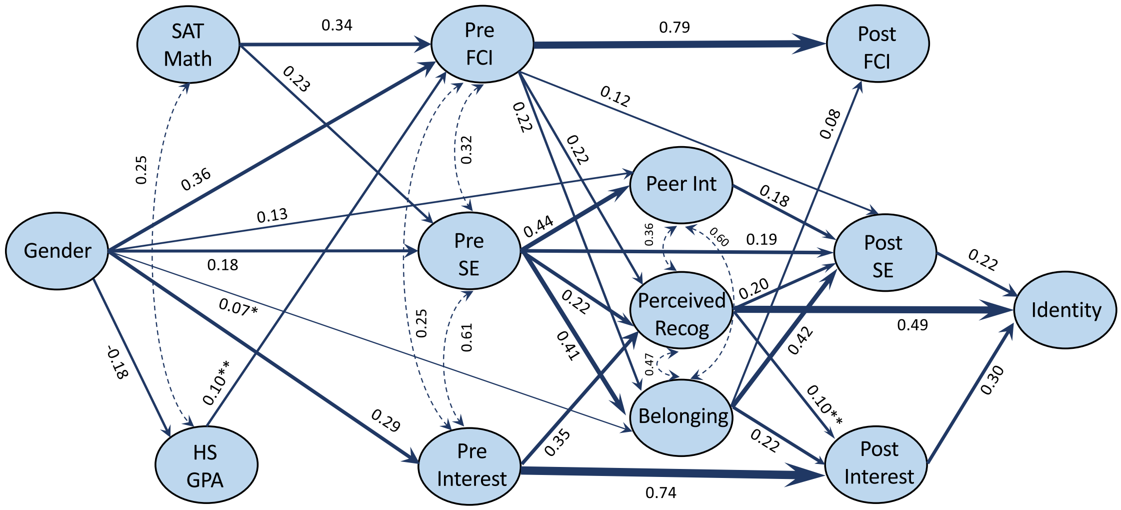

FIG. 5. Schematic diagram of the path analysis part of the structural equation modeling (model 4) between gender and overall physics identity through SAT Math scores, high school GPA (HS GPA), and FCI scores as well as peer interaction (Int), perceived recognition (Recog), sense of belonging, self-efficacy (SE), and interest. The solid lines represent regression paths and the dashed lines represent residual covariances. The regression line thickness corresponds to the magnitude of β value (standardized regression coefficient) with $0.01 \leq p < 0.05$ indicated by * and $0.001 \leq p < 0.01$ indicated by **. All the other regression lines show relations with $p < 0.001$.

largest predictor. Postinterest is predicted by perceived recognition and sense of belonging, and post-FCI is predicted by sense of belonging. Similar to models 1–3, students' overall physics identity is directly predicted by perceived recognition, post-self-efficacy, and postinterest, and perceived recognition is the largest predictor.

Although Tables III and IV show that there were large gender differences disadvantaging women in students' FCI scores, self-efficacy, interest, and overall physics identity at the end of the course, we note that gender does not directly predict these constructs in any of the models discussed. Thus, our results reveal that the gender differences in these outcome constructs were mediated through the different constructs of the model including components of students' perception of the inclusiveness of the learning environment.

5. Direct and indirect paths in model 4

Model 4 shows that the three components of students' perception of the inclusiveness of the learning environment not only directly predict the outcome constructs but also mediate the indirect effect of premotivational beliefs and FCI scores on postmotivational beliefs and FCI scores. To summarize how the outcome constructs were predicted by different predictors through both direct and indirect paths, we calculated the regression coefficient for each path in model 4. The results are shown in Table VI. For example, there are three different indirect paths from pre-self-efficacy to post-self-efficacy mediated through peer interaction, perceived recognition, and sense of belonging, respectively. The indirect effect of pre-self-efficacy on post-self-efficacy can be calculated by adding these three paths together ($\beta = 0.44 \times 0.18 + 0.22 \times 0.20 + 0.41 \times 0.42 = 0.30$), which is larger than the direct effect of pre-self-efficacy on

TABLE VI. Regression coefficients (β) of direct and indirect paths for the four outcome constructs predicted by various predictors in model 4.

Outcome	Predictor	Direct	Indirect	Total
Post FCI	SAT Math	0.00	0.28	0.28
	High school GPA	0.00	0.08	0.08
	Pre-FCI	0.79	0.02	0.81
	Pre-self-efficacy	0.00	0.03	0.03
	Pre-interest	0.00	0.00	0.00
	Peer interaction	0.00	0.00	0.00
	Perceived recognition	0.00	0.00	0.00
	Belonging	0.08	0.00	0.08
Post self-efficacy	SAT Math	0.00	0.20	0.20
	High school GPA	0.00	0.03	0.03
	Pre-FCI	0.12	0.14	0.26
	Pre-self-efficacy	0.19	0.30	0.49
	Pre-interest	0.00	0.07	0.07
	Peer interaction	0.18	0.00	0.18
	Perceived recognition	0.20	0.00	0.20
	Belonging	0.42	0.00	0.42
Postinterest	SAT Math	0.00	0.05	0.05
	High school GPA	0.00	0.01	0.01
	Pre-FCI	0.00	0.07	0.07
	Pre-self-efficacy	0.00	0.11	0.11
	Pre-interest	0.74	0.04	0.78
	Peer interaction	0.00	0.00	0.00
	Perceived recognition	0.10	0.00	0.10
	Belonging	0.22	0.00	0.22
Overall physics identity	SAT Math	0.00	0.12	0.12
	High school GPA	0.00	0.02	0.02
	Pre-FCI	0.00	0.19	0.19
	Pre-self-efficacy	0.00	0.25	0.25
	Pre-interest	0.00	0.42	0.42
	Peer Interaction	0.00	0.04	0.04
	Perceived recognition	0.49	0.07	0.56
	Belonging	0.00	0.16	0.16

post-self-efficacy ($\beta = 0.19$). We note that the direct effect of students' sense of belonging on post-self-efficacy ($\beta = 0.42$) is almost the same as the total effect of pre-self-efficacy on post-self-efficacy ($\beta = 0.49$). In addition, we found that even though students' post-interest is mainly predicted by their pre-interest, it is also predicted by their sense of belonging ($\beta = 0.22$) and perceived recognition ($\beta = 0.10$). Similarly, even though post-FCI is mainly predicted by pre-FCI, it is also predicted by sense of belonging with $\beta = 0.08$. Even though Fig. 5 shows that perceived recognition is the only inclusiveness of learning environment construct that predicts overall physics identity, Table VI shows that students' sense of belonging also indirectly predicts their overall physics identity with $\beta = 0.16$.

6. Comparison between models

To further understand the role played by each inclusiveness of learning environment construct in predicting the outcome constructs, we compared the four SEM models discussed earlier. As shown in Table VII, first, we summarize the regression coefficients from perceived recognition, peer interaction, and sense of belonging to the outcome constructs in the four models. Then, we calculated the coefficients of determination R^2 (fraction of variance explained) for each outcome construct in the four models. Finally, we summarize the fit indices for each model. The model fit is good if the fit parameters are above certain thresholds. In particular, CFI > 0.9, TLI > 0.9, RMSEA < 0.08, and SRMR < 0.08 are considered as acceptable and RMSEA < 0.06 and SRMR < 0.06 are considered as a good fit [172]. As shown in Table VII, the four models have very comparable fit indices and they all fit the data very well.

By comparing the regression coefficients from different inclusiveness of learning environment constructs to post-FCI in the four models, we find that perceived recognition is a direct predictor of post-FCI in model 1, while this effect is no longer statistically significant in models 2–4 after controlling for peer interaction or sense of belonging. On the other hand, we note that the direct effect from sense of belonging to post-FCI is statistically significant even after controlling for both perceived recognition and peer interaction in model 4. In addition, Table VII shows that although all three inclusiveness of learning environment factors are significant predictors of students' post-self-efficacy, sense of belonging is always the largest predictor compared when it is included in the model. We note that in all four models, perceived recognition is a direct predictor of students' overall physics identity, while the direct effects of sense of belonging and peer interaction are not statistically significant.

TABLE VII. Summary of the regression coefficients from learning environment components to outcome constructs, coefficient of determination (R^2) for various outcome constructs, and model fit indices for different models with different combinations of perceived recognition (Recog), peer interaction, and sense of belonging (Bel) as predictors. All regression coefficients shown are statistically significant. ns represents not statistically significant. All R^2 values are significant with p values <0.001.

Regression coefficients from learning environment components to outcome constructs								
	Model 1	Model 2		Model 3		Model 4		
	Recog	Recog	Peer	Recog	Bel	Recog	Peer	Bel
Post-FCI	0.06	ns	ns	ns	0.07	ns	ns	0.08
Post-self-efficacy	0.48	0.32	0.38	0.22	0.52	0.20	0.18	0.42
Post-interest	0.22	0.14	0.18	0.11	0.20	0.10	ns	0.22
Overall physics identity	0.52	0.51	ns	0.50	ns	0.49	ns	ns

Coefficient of determination (R^2) for different outcome constructs				
	Model 1	Model 2	Model 3	Model 4
Post-FCI	0.68	0.68	0.68	0.68
Post-self-efficacy	0.60	0.70	0.75	0.77
Post-interest	0.82	0.82	0.82	0.82
Overall physics identity	0.79	0.80	0.79	0.80

Fit indices				
	Model 1	Model 2	Model 3	Model 4
CFI	0.987	0.988	0.979	0.982
TLI	0.986	0.987	0.977	0.981
RMSEA	0.051	0.047	0.054	0.049
SRMR	0.053	0.051	0.053	0.051

Next, we compared the coefficients of determination R^2 (fraction of variance explained) for each outcome construct in the four models. We find that in all four models, R^2 values of outcome constructs are reasonably high, which means that our models have explained much of the variance in them. In particular, we note that the R^2 values of post-FCI, post-self-efficacy, and post-interest are almost the same across different models. That means that each model can explain 68% of the variance in post-FCI, around 82% of the variance in postinterest, and 80% of the variance in overall physics identity. On the other hand, different models explain different amount of variance in post-self-efficacy. In particular, the models including sense of belonging always explain more variance in post-self-efficacy than the models without sense of belonging do. These results are consistent with the finding discussed earlier that sense of belonging is the major predictor of post-self-efficacy. Table VII shows that model 4 has the largest R^2 value for post-self-efficacy compared with the other three models, which means that model 4 can best explain the variance in post-self-efficacy. Considering that there are only very small differences between different models' fit indices, we believe that model 4, which includes all three inclusiveness of learning environment constructs, is most productive.

VII. SUMMARY AND DISCUSSION

In this study, we focused on students' physics motivational beliefs and FCI scores in a college calculus-based introductory physics course at a large public research university. We studied how students' perception of the inclusiveness of the learning environment—including peer interaction, perceived recognition, and sense of belonging—predicts students' motivational beliefs and FCI scores at the end of the course after controlling for their gender, high school performance, and motivational beliefs and FCI scores at the beginning of the course.

In response to RQ2, our results show that the inclusiveness of the learning environment statistically significantly predicts students' motivational beliefs and FCI scores at the end of the course. There was no statistically significant gender difference in the relationship between any two constructs in the models (RQ3). Moreover, even though we found that there are statistically significant gender differences disadvantaging women in self-efficacy, interest, overall physics identity and FCI scores at the end of the course (RQ1), gender only directly predicts the controlled factors and inclusiveness of learning environment constructs and does not directly predict any outcome constructs (RQ4). This implies that the gender differences in these learning outcomes were mediated by students' perception of the inclusiveness of the learning environment. Thus, in addition to being driven by prior differences, which often result from inequities including societal stereotypes and biases about who belongs in physics and lack of role models, students' self-efficacy, interest, overall physics identity and FCI scores are also influenced by their perception of the inclusiveness of learning environment [182]. Furthermore, our results show that the current learning environment is not helping to reduce the gender difference, and instead, the gender difference in students' self-efficacy increased by the end of the course (RQ1). We note that, in the current learning environment, female students also reported less benefit from peer interaction, felt a lower sense of belonging and felt less recognized as a physics person than male students, which may all contribute to the gender differences in students' learning outcomes at the end of the course. For example, in a male-dominated classroom environment, a woman may experience a lower level of sense of belonging and higher level of anxiety with lower self-efficacy than men [24]. In addition, nonsupportive instructional pedagogies, lack of recognition from instructors and TAs and lack of positive interactions with peers can further decrease women's self-efficacy in physics. Thus, the instructor's focus on equity and inclusion, and approaches to recognizing students in poorly gender-balanced classrooms, become even more vital in supporting women's self-efficacy and promoting learning for all students in the classroom [124].

Our findings also suggest that students' perception of the inclusiveness of the learning environment plays a very important role in explaining their motivational beliefs and performance at the end of the course. In response to **RQ5**, we found that perceived recognition contributed most to predicting overall physics identity, and sense of belonging contributed most to predicting self-efficacy. We note that even though peer interaction has a smaller direct effect on the outcome constructs compared with perceived recognition and sense of belonging, this does not mean that effective peer interaction is not important. Many instructors may not know how to implement strategies to improve students' sense of belonging. The correlation between peer interaction and the other two inclusiveness of learning environment constructs suggests a possibility that students' sense of belonging and perceived recognition may possibly be shaped by helping students interact meaningfully with peers (which in turn can improve student outcomes). For example, prior studies have hinted at the fact that the learning environment is an interconnected ecological system rather than the simple sum of its parts [183]. Moreover, we note that the model including all three inclusiveness of learning environment constructs can best explain the variance in outcome constructs compared with the other three models studied. Therefore, we believe the model including all three inclusiveness of learning environment constructs is most productive (RQ6).

By comparing students' responses to the survey in pre and post, we found that both male and female students' self-efficacy and interest statistically significantly dropped from pre to post. And female students' motivational beliefs dropped even more than male students', which may partially explain that the gender difference in students' self-efficacy increased by the end of the course. These results indicate that the current learning environment is not helping students improve their physics motivational beliefs and, on the contrary, contributes to decreasing them in such a way that the gender gap increases.

Therefore, instructors must make intentional efforts to help students improve their physics motivational beliefs and performance within the equity of parity framework discussed earlier (i.e., regardless of the initial value at the beginning of the course, instructors should strive to ensure that at the end of the course, all demographic groups have similar high levels of motivational beliefs and performance). As noted, the perception of the inclusiveness of the learning environment directly predicts students' motivational outcomes and post-FCI scores, so it is reasonable to expect that a more inclusive learning environment will help. Instructors should strive to reduce the effects of prior preparation and prior motivational beliefs so that all students can equally benefit from the learning environment. If we could eliminate the gender difference in sense of belonging, perceived recognition and peer interaction by creating a learning environment, in which all students feel safe to engage in collaboration and discussions with peers and instructor, and provide appropriate scaffolding support commensurate with students' prior knowledge, the gender difference in students' motivational beliefs and FCI scores may also decrease.

Evidence-based instructional strategies may be helpful for instructors to improve the inclusiveness of the learning environment and support traditionally marginalized students such as women in physics. For example, instructors can provide students with opportunities to engage in different types of interaction, such as setting up study groups or assigning collaborative tasks [184]. However, instructors need to keep in mind how societal stereotypes and biases about who belongs in physics and can excel in it impact the stereotyped groups and avoid letting a small group of students dominate the discussion so that all students' voices can be heard and valued. Another stereotype about physics is that it requires a natural ability to excel [185,186]. Studies have shown that the idea of ability being fixed and unchangeable can increase students' concerns about belonging, especially for students from traditionally marginalized groups such as women in physics who have few role models [187,188]. Thus, it is critical to build a learning environment that emphasizes that abilities are malleable and can be changed through deliberate practice and effort [189]. Instructors can also show students nonstereotypical role models from diverse demographic groups, personalities, and interest in different contexts since this has been shown to increase students' sense of belonging [190,191]. In addition, instructors can explicitly recognize students by directly acknowledging their work and expressing faith in their ability, and they can also implicitly recognize students by valuing students' opinions and assigning a leadership position or a challenging task to students in small groups that makes them feel valued [192]. However, instructors should be careful not to give unintended messages to students, e.g., praising some students for brilliance or intelligence as opposed to their effort since it may convey to other students that they do not have what is required to excel in physics [185,186].

VIII. LIMITATIONS AND FUTURE DIRECTIONS

In this study, we discussed how students' perception of the inclusiveness of the learning environment predicts female and male students' motivational beliefs and FCI scores in the introductory calculus-based physics course. This study is a field study, in which we did not have experimental manipulation or intervention with random assigned control groups to investigate the effect of students' perception of inclusiveness of the learning environment. We did use hierarchical linear modeling (HLM) to test the instructor level effects on students' motivational beliefs and the results show that the instructor level effects can be ignored [193]. In future, it would be valuable to conduct controlled studies to further investigate the role played by inclusiveness of learning environment. In addition, this study is based on students' self-reported responses to a survey with Likert scale response options. It would be helpful to interview more students to get a deeper qualitative understanding of what they experienced during the learning process in the course, and how their experiences affected their motivational beliefs and learning outcomes.

In this study, we used a single item as a holistic measure of students' overall physics identity, which may not capture the full complexity of physics identity. Even though this item is commonly used in studies involving physics identity [65,80–82], it would be helpful in future studies to develop more survey items for physics identity construct. In addition, in this study, we focused on students' perception of the inclusiveness of the learning environment, which could be different from the perceptions of instructors or TAs or a third party who observes the course. Future studies can investigate the roles played by the perceptions of different groups of people in predicting students' course outcomes, which may be also helpful in developing a better understanding of how to build an inclusive learning environment.

This study was conducted in a traditionally taught introductory calculus-based physics course. It would be interesting to investigate students' perception of the inclusiveness of the learning environment 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 focus on equity and inclusion and compare the results with those of the current study. Future studies can also investigate the inclusiveness of the learning environment in other courses, such as algebra-based physics courses, where women are often the majority group, or advanced physics courses beyond the first year, which are typically taken by physics majors. In the future studies, we also intend to carry out similar investigations accounting for intersectional perspectives, e.g., with female and male students from different ethnic or racial groups and how their perceptions of the inclusiveness of learning environment predict their course outcomes. In addition, our study was conducted in a large public research university in the U.S. Similar studies in different types of institutions such as small colleges and universities in the U.S. and in other countries would also be helpful for developing a deeper understanding of the relationships between students' perceptions of the inclusiveness of learning environment and their course outcomes.

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 A: DATA DISTRIBUTION

The figure below (Fig. 6) presents the distributions of students' high school GPA, SAT math scores, pre-FCI scores and post-FCI scores.

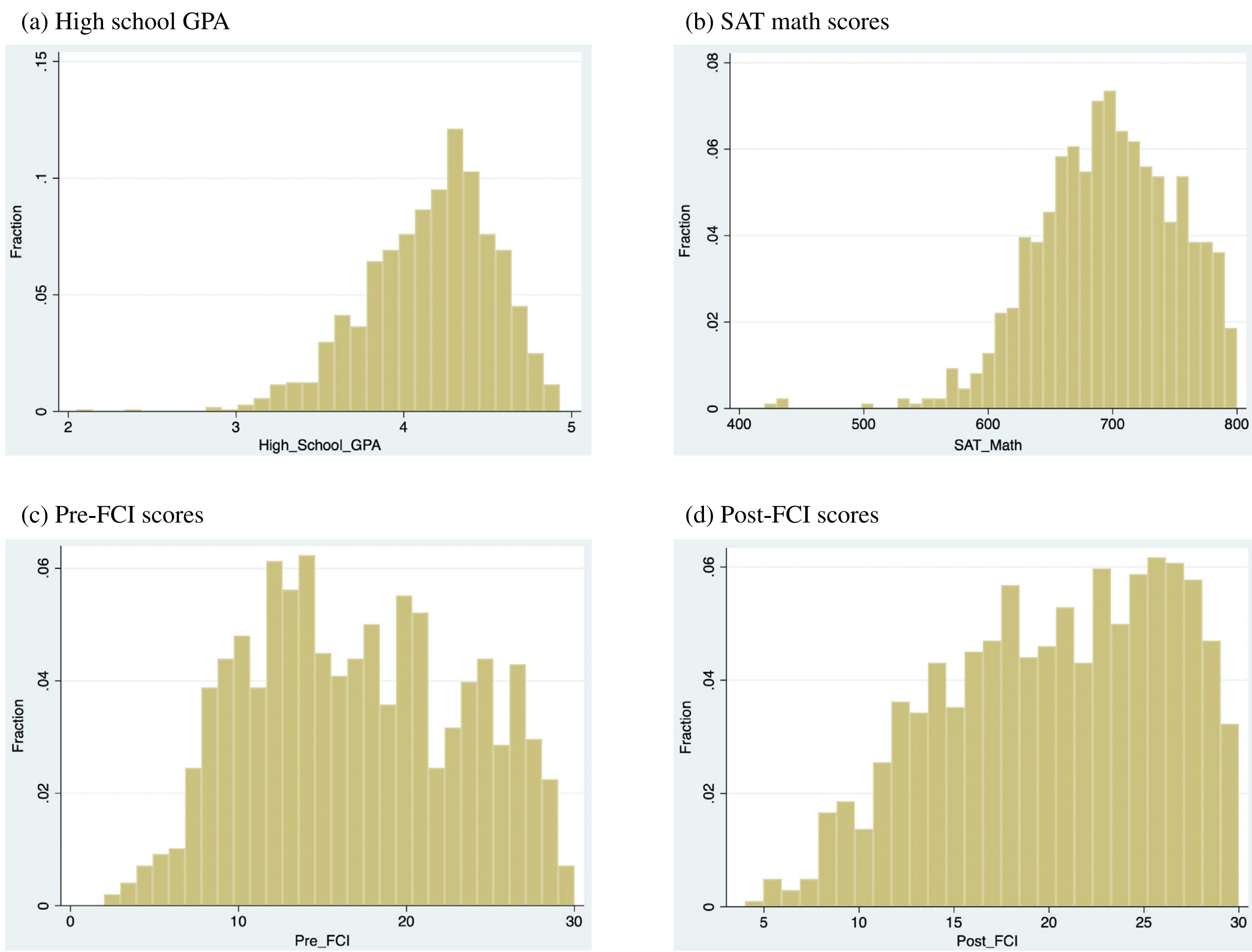

FIG. 6. Graphs of the distributions of (a) high school GPA, (b) SAT math score, (c) pre-FCI scores, and post-FCI scores.

APPENDIX B: PERCENTAGES OF STUDENTS WHO SELECTED EACH CHOICE FOR EACH SURVEY ITEM

In the main text, we discussed how students' motivational beliefs change from the beginning (pre) to the end (post) of the course by comparing their average scores on the pre- and postmotivational constructs. Here, we present the percentages of female and male students who selected each answer choice from a Likert scale for each survey item (Tables VIII–XI). The survey items for sense of belonging were scored on a 5-point Likert scale, while the survey items for all the other motivational constructs were scored on a 4-point Likert scale. For all survey items, higher scores indicate greater levels of motivational beliefs.

TABLE VIII. Percentages of female and male students who selected each choice from a 4-point Likert scale for each survey item of self-efficacy (SE) in the pre- and postsurvey, which have the response scale: 1 = NO!, 2 = no, 3 = yes, and 4 = YES!.

		Pre				Post			
	Survey items	1	2	3	4	1	2	3	4
Female	SE1	7%	29%	54%	10%	9%	30%	55%	6%
	SE2	1%	11%	75%	12%	4%	16%	71%	8%
	SE3	1%	4%	64%	31%	5%	28%	54%	13%
	SE4	1%	10%	70%	19%	5%	24%	61%	10%
Male	SE1	3%	25%	60%	12%	4%	22%	63%	11%
	SE2	1%	8%	71%	21%	1%	10%	70%	19%
	SE3	1%	3%	55%	41%	2%	14%	56%	28%
	SE4	0%	7%	69%	24%	2%	16%	66%	16%

TABLE IX. Percentages of female and male students who selected each choice from a 4-point Likert scale for each survey item of interest in the pre- and post-survey. Interest1 has the response scale: 1 = Never, 2 = Once a month, 3 = Once a week, 4 = Every day". Interest2 has the response scale: 1 = Very boring, 2 = boring, 3 = interesting, 4 = Very interesting. The other two items have the response scale: 1 = NO!, 2 = no, 3 = yes, and 4 = YES!.

		Pre				Post			
	Survey items	1	2	3	4	1	2	3	4
Female	Interest1	8%	36%	41%	15%	8%	20%	47%	25%
	Interest2	3%	14%	64%	20%	6%	19%	62%	13%
	Interest3	2%	26%	55%	17%	6%	41%	42%	10%
	Interest4	1%	23%	57%	19%	7%	30%	50%	12%
Male	Interest1	4%	22%	43%	31%	3%	12%	41%	44%
	Interest2	1%	5%	62%	32%	2%	8%	61%	28%
	Interest3	1%	12%	55%	32%	3%	22%	49%	25%
	Interest4	1%	13%	64%	23%	3%	20%	52%	25%

TABLE X. Percentages of female and male students who selected each choice from a 4-point Likert scale for each survey item of peer interaction, perceived recognition, and physics identity. All items have the response scale: 1 = strongly disagree, 2 = disagree, 3 = agree, and 4 = strongly agree.

	Female				Male			
Survey items	1	2	3	4	1	2	3	4
Peer1	6%	17%	60%	17%	2%	17%	55%	26%
Peer2	8%	26%	57%	10%	3%	17%	58%	22%
Peer3	8%	29%	53%	10%	3%	20%	55%	22%
Peer4	8%	33%	50%	8%	4%	22%	56%	19%
Recognition1	18%	38%	35%	8%	9%	30%	44%	17%
Recognition2	16%	40%	36%	8%	9%	32%	44%	15%
Recognition3	21%	48%	28%	2%	12%	42%	40%	6%
Identity1	21%	45%	28%	6%	9%	35%	42%	15%

TABLE XI. Percentages of female and male students who selected each choice from a 5-point Likert scale for each survey item of sense of belonging. All items have the response scale: 1 = not at all true, 2 = a little true, 3 = somewhat true, 4 = mostly true, and 5 = completely true.

	Female					Male				
Survey items	1	2	3	4	5	1	2	3	4	5
Belonging1	10%	15%	30%	30%	14%	4%	9%	28%	35%	24%
Belonging2	4%	8%	21%	37%	31%	2%	5%	11%	34%	48%
Belonging3	11%	22%	30%	29%	8%	5%	16%	31%	33%	15%
Belonging4	6%	18%	30%	35%	12%	4%	12%	32%	36%	17%
Belonging5	8%	14%	28%	24%	26%	5%	9%	18%	30%	39%

As shown in Table VIII, for both female and male students, the percentages of students who selected 4 decreased from pre to post for all self-efficacy items, while the percentages of students who selected 1 or 2 mostly increased. Table IX shows similar shifts in students' responses to the survey items under interest. These results are consistent with the descriptive statics shown in Table III, which show that both male and female students' self-efficacy and interest statistically significantly decreased from pre to post.

In addition, by comparing percentages of female and male students who selected each answer choice, we found that for most survey items, the percentages of female students who selected 1 or 2 were larger than those of male students, while the percentages of female students who selected 4 (for sense of belonging is 5) were smaller than those of male students. These findings are also consistent with Tables III and IV showing that there were statistically significant gender differences in all motivational constructs studied.

APPENDIX C: MODERATION ANALYSIS

We conducted a moderation analysis to test whether gender moderates the relationship between any two constructs in the models (i.e., do the strength of relationships given by the standardized regression coefficients between any two constructs in the models differ for women and men). We used the R [163] software package "lavaan" to conduct multigroup SEM. We initially tested for measurement invariance. In other words, we looked at 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 where 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 intercept 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 $(\Delta\chi^2) = 25.001$, degree of freedom difference $(\Delta\text{dof}) = 21$, and nonsignificant $p = 0.2471$]. 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 ($\Delta\chi^2 = 27.924$, $\Delta\text{dof} = 21 = 21$, $p = 0.1423$) or the freely estimated model ($\Delta\chi^2 = 52.925$, $\Delta\text{dof} = 42$, $p = 0.1203$), so strong measurement invariance holds. Finally, to test for "strict" measurement invariance we fixed factor loadings, intercepts, and residual variances to equality. In this step, there was a statistically significant difference from the scalar measurement model ($\Delta\chi^2 = 60.908$, $\Delta\text{dof} = 27$, $p < 0.001$), therefore "strict invariance" did not hold. However, strict invariance is unlikely to hold in most situations. Therefore, since strong measurement invariance holds for this model, we continued on to perform other group comparisons.

Next, 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 where regression pathways are equal for men and women. The model fit parameters for this case were acceptable (RMSEA = 0.056, SRMR = 0.058, CFI = 0.914, TLI = 0.910). The multigroup SEM results suggest that regression pathways among the constructs do not have differences across gender when we compared to the freely estimated model ($\Delta\chi^2 = 93.438$, $\Delta\text{dof} = 74$, $p = 0.063$) or to the scalar model ($\Delta\chi^2 = 40.513$, $\Delta\text{dof} = 32$, $p = 0.1437$).

[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. Kalender Yasemin, 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. Innov. Sci. Math. Educ. (formerly CAL-laborate International) **19**, 1 (2011).

[19] J. Blue, A. L. Traxler, and X. C. Cid, Gender matters, Phys. Today **71**, 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] 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. Policy Rev. **33**, 304 (2013).

[24] 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. Engin. Educ. **84**, 151 (1995).

[25] 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).

[26] A. W. *et al.*, Social emotionaly and personality development, *Handbook of Child Psychology* (Wiley, New York, 2006), Vol. 3.

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

[34] J. C. Blickenstaff, Women and science careers: Leaky pipeline or gender filter?, Gender Educ. **17**, 369 (2005).
[35] D. H. Schunk and F. Pajares, The Development of Academic Self-Efficacy, in *Development of Achievement Motivation: A Volume in the Educational Psychology Series*, edited by A. Wigfield and J. S. Eccles (Academic Press, San Diego, 2002), p. 15.
[36] E. Seymour, Tracking the processes of change in US undergraduate education in science, mathematics, engineering, and technology, Sci. Educ. **86**, 79 (2002).
[37] S. G. Brainard and L. Carlin, A six-year longitudinal study of undergraduate women in engineering and science, J. Engin. Educ. **87**, 369 (1998).
[38] S. J. Correll, Gender and the career choice process: The role of biased self-assessments, Am. J. Sociology **106**, 1691 (2001).
[39] S. J. Correll, Constraints into preferences: Gender, status, and emerging career aspirations, Am. Sociol. Rev. **69**, 93 (2004).
[40] 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).
[41] E. T. Pascarella and P. T. Terenzini, *How College Affects Students: Findings and Insights from Twenty Years of Research* (Jossey-Bass Inc., San Francisco, CA, 1991).
[42] 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).
[43] 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).
[44] 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).
[45] 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).
[46] S. Cheryan and V. C. Plaut, Explaining underrepresentation: A theory of precluded interest, Sex Roles **63**, 475 (2010).
[47] 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).
[48] K. G. Talley and A. M. Ortiz, Women's interest development and motivations to persist as college students in STEM: A mixed methods analysis of views and voices from a Hispanic-serving institution, Int. J. STEM Educ. **4**, 5 (2017).
[49] A. Wigfield and J. S. Eccles, The development of achievement task values: A theoretical analysis, Dev. Policy Rev. **12**, 265 (1992).
[50] A. Wigfield and J. S. Eccles, Expectancy–value theory of achievement motivation, Contemp. Educ. Psychol. **25**, 68 (2000).
[51] A. Bandura, Self-efficacy, in *Encyclopedia of Psychology*, 2nd ed., edited by R. J. Corsini (Wiley, New York, 1994), Vol. 3, p. 368.
[52] A. Bandura, Social cognitive theory of self-regulation, Organ. Behav. Hum. Decis. Process. **50**, 248 (1991).
[53] S. L. Britner and F. Pajares, Sources of science self-efficacy beliefs of middle school students, J. Res. Sci. Teach. **43**, 485 (2006).
[54] Z. D. Kirbulut and E. Uzuntiryaki-Kondakci, Examining the mediating effect of science self-efficacy on the relationship between metavariables and science achievement, Int. J. Sci. Educ. **41**, 995 (2019).
[55] J. M. Harackiewicz, K. E. Barron, J. M. Tauer, and A. J. Elliot, Predicting success in college: A longitudinal study of achievement goals and ability measures as predictors of interest and performance from freshman year through graduation, J. Educ. Psychol. **94**, 562 (2002).
[56] S. Hidi, Interest: A unique motivational variable, Educ. Res. Rev. **1**, 69 (2006).
[57] K. A. Renninger and S. Hidi, Revisiting the conceptualization, measurement, and generation of interest, Educ. Psychol. **46**, 168 (2011).
[58] A. Wigfield, J. S. Eccles, U. Schiefele, R. W. Roeser, and P. Davis-Kean, Development of achievement motivation, in *Handbook of Child Psychology: Social, Emotional, and Personality Development*, Vol. 3, 6th ed., edited by N. Eisenberg, W. Damon, and R. M. Lerner (John Wiley & Sons Inc, Hoboken, NJ, US, 2006), p. 933.
[59] H.-S. Lin, Z.-R. Hong, and Y.-C. Chen, Exploring the development of college students' situational interest in learning science, Int. J. Sci. Educ. **35**, 2152 (2013).
[60] E. B. Witherspoon, C. D. Schunn, R. M. Higashi, and E. C. Baehr, Gender, interest, and prior experience shape opportunities to learn programming in robotics competitions, Int. J. STEM Educ. **3**, 18 (2016).
[61] P. Häussler and L. Hoffmann, An intervention study to enhance girls' interest, self-concept, and achievement in physics classes, J. Res. Sci. Teach. **39**, 870 (2002).
[62] Z. Hazari, P. M. Sadler, and G. Sonnert, The science identity of college students: Exploring the intersection of gender, race, and ethnicity, J. Coll. Sci. Teach. **42**, 82 (2013), https://www.jstor.org/stable/43631586.
[63] C. Monsalve, Z. Hazari, D. McPadden, G. Sonnert, and P. M. Sadler, Examining the relationship between career outcome expectations and physics identity, in *Proceedings of PER Conf. 2016, Sacramento, CA*, 2016, 10.1119/perc.2016.pr.052.
[64] G. Potvin and Z. Hazari, The development and measurement of identity across the physical sciences, in *Proceedings of PER Conf. 2013,Portland, OR*, 2013, 10.1119/perc.2013.pr.058.
[65] 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. Engin. Educ. **105**, 312 (2016).

[66] A. Y. Kim and G. M. Sinatra, Science identity development: An interactionist approach, Int. J. STEM Educ. **5**, 51 (2018).

[67] K. Atkins, B. M. Dougan, M. S. Dromgold-Sermen, H. Potter, V. Sathy, and A. Panter, "Looking at Myself in the Future": How mentoring shapes scientific identity for STEM students from underrepresented groups, Int. J. STEM Educ. **7**, 42 (2020).

[68] A. Singer, G. Montgomery, and S. Schmoll, How to foster the formation of STEM identity: Studying diversity in an authentic learning environment, Int. J. STEM Educ. **7**, 57 (2020).

[69] P. Vincent-Ruz and C. D. Schunn, The nature of science identity and its role as the driver of student choices, Int. J. STEM Educ. **5**, 48 (2018).

[70] R. M. Lock, Z. Hazari, and G. Potvin, Physics career intentions: The effect of physics identity, math identity, and gender, AIP Conf. Proc. **1513**, 262 (2013).

[71] Z. Y. Kalender, Ph.D. thesis, University of Pittsburgh, Pittsburgh, PA, 2019.

[72] L. Archer, J. Moote, B. Francis, J. DeWitt, and L. Yeomans, The "exceptional" physics girl: A sociological analysis of multimethod data from young women aged 10–16 to explore gendered patterns of post-16 participation, Am. Educ. Res. J. **54**, 88 (2017).

[73] A. T. Danielsson, Exploring woman university physics students 'doing gender' and 'doing physics', Gender Educ. **24**, 25 (2012).

[74] A. J. Gonsalves, Exploring how gender figures the identity trajectories of two doctoral students in observational astrophysics, Phys. Rev. Phys. Educ. Res. **14**, 010146 (2018).

[75] N. M. Hewitt and E. Seymour, A long, discouraging climb, ASEE Prism **1**, 24 (1992).

[76] E. Seymour, N. M. Hewitt, and C. M. Friend, *Talking About Leaving: Why Undergraduates Leave the Sciences* (Westview Press, Boulder, CO, 1997), Vol. 12.

[77] G. Potvin and Z. Hazari, Student evaluations of physics teachers: On the stability and persistence of gender bias, Phys. Rev. Phys. Educ. Res. **12**, 020107 (2016).

[78] J. M. Youngblut and G. R. Casper, Focus on psychometrics single-item indicators in nursing research, Res. Nursing Health **16**, 459 (1993).

[79] P. A. Patrician, Single-item graphic representational scales, Nursing Res. **53**, 347 (2004).

[80] A. Godwin, G. Potvin, Z. Hazari, and R. Lock, Understanding engineering identity through structural equation modeling, in *Proceedings of the 2013 IEEE Frontiers in Education Conference (FIE)*, (IEEE, Piscataway, NJ, 2013), p. 50.

[81] 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).

[82] 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).

[83] J. Wang, Q. Li, and Y. Luo, Physics identity of Chinese students before and after Gaokao: The effect of high-stake testing, Res. Sci. Educ. **52**, 675 (2022).

[84] H. Cheng, G. Potvin, R. Khatri, L. H. Kramer, R. M. Lock, and Z. Hazari, Examining physics identity development through two high school interventions, in *Proceedings of PER Conf. 2018, Washington, DC*, 10.1119/perc.2018.pr.Cheng.

[85] A. Prybutok, A. Patrick, M. Borrego, C. C. Seepersad, and M. Kirisits, Cross-sectional survey study of undergraduate engineering identity, in *Proceedings of the American Society for Engineering Education Annual Conference, New Orleans, Louisiana* (2016).

[86] J. D. Cribbs, Z. Hazari, G. Sonnert, and P. M. Sadler, Establishing an explanatory model for mathematics identity, Child Development **86**, 1048 (2015).

[87] J. Cribbs, Z. Hazari, P. M. Sadler, and G. Sonnert, Development of an explanatory framework for mathematics identity, in *Proceedings of the Psychology of Mathematics Education-North American (PME-NA) Chapter Conference* (2012).

[88] K. Miller, J. Schell, A. Ho, B. Lukoff, and E. Mazur, Response switching and self-efficacy in Peer Instruction classrooms, Phys. Rev. ST Phys. Educ. Res. **11**, 010104 (2015).

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

[90] K. A. Robinson, T. Perez, A. K. Nuttall, C. J. Roseth, and L. Linnenbrink-Garcia, From science student to scientist: Predictors and outcomes of heterogeneous science identity trajectories in college, Develop. Psychol. **54**, 1977 (2018).

[91] M. M. Chemers, E. L. Zurbriggen, M. Syed, B. K. Goza, and S. Bearman, The role of efficacy and identity in science career commitment among underrepresented minority students, J. Soc. Issues **67**, 469 (2011).

[92] A. Bandura, W. Freeman, and R. Lightsey, Self-efficacy: The exercise of control, J. Cogn. Psychotherapy **13**, 158 (1999).

[93] A. Bandura, Self-efficacy: Toward a unifying theory of behavioral change, Psychol. Rev. **84**, 191 (1977).

[94] B. M. Hagerty, R. A. Williams, J. C. Coyne, and M. R. Early, Sense of belonging and indicators of social and psychological functioning, Arch. Psychiatr. Nurs. **10**, 235 (1996).

[95] C. Choenarom, R. A. Williams, and B. M. Hagerty, The role of sense of belonging and social support on stress and depression in individuals with depression, Arch. Psychiatr. Nurs. **19**, 18 (2005).

[96] S. Hidi and K. A. Renninger, The four-phase model of interest development, Educ. Psychol. **41**, 111 (2006).

[97] K. A. Renninger and S. Hidi, Student interest and achievement: Developmental issues raised by a case study, *Development of Achievement Motivation* (Elsevier, New York, 2002), p. 173.

[98] C. A. Wolters, Self-regulated learning and college students' regulation of motivation, J. Educ. Psychol. **90**, 224 (1998).

[99] S. Hidi, J. Weiss, D. Berndorff, and J. Nolan, The role of gender, instruction and a cooperative learning technique in science education across formal and informal settings, in *Interest and learning: Proceedings of the Seeon conference on interest and gender, IPN Kiel, Germany* (IPN, Kiel, Germany, 1998), p. 215.
[100] M. Mitchell, Situational interest: Its multifaceted structure in the secondary school mathematics classroom, J. Educ. Psychol. **85**, 424 (1993).
[101] K. Renninger and W. Shumar, *Community Building with and for Teachers at the Math Forum* (Cambridge University Press, Cambridge, England, 2002), 10.1017/CBO9780511606373.008.
[102] L. A. Sosniak, The tortoise, the hare, and the development of talent, *Encouraging the Development of Exceptional Abilities and Talents*, edited by M. J. A. Howe (The British Psychological Society, Leicester, 1990), pp. 149–164.
[103] T. Nouh, S. Anil, A. Alanazi, W. Al-Shehri, N. Alfaisal, B. Alfaris, and E. Alamer, Assessing correlation between students' perception of the learning environment and their academic performance, JPMA The Journal of the Pakistan Medical Association **66**, 1616 (2016).
[104] M. C. Hill and K. K. Epps, Does physical classroom environment affect student performance, student satisfaction, and student evaluation of teaching in the college environment?, *Allied Academies International Conference. Academy of Educational Leadership. Proceedings* (Jordan Whitney Enterprises, Inc., New Orleans, 2009), p. 15.
[105] L. McCullough, Gender differences in student responses to physics conceptual questions based on question context, ASQ Advancing the STEM Agenda in Education (2011).
[106] 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, https://web.physics.indiana.edu/hake/PERC2002h-Hake.pdf.
[107] J. Docktor and K. Heller, Gender differences in both force concept inventory and introductory physics performance, AIP Conf. Proc. **1064**, 15 (2008).
[108] 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).
[109] 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).
[110] D. Hestenes, M. Wells, and G. Swackhamer, Force Concept Inventory, Phys. Teach. **30**, 141 (1992).
[111] L. McCullough, Gender, context, and physics assessment, J. Int. Women's Studies **5**, 20 (2004), https://vc.bridgew.edu/jiws/vol5/iss4/2.
[112] 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).
[113] 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).
[114] 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).
[115] V. P. Coletta, J. A. Phillips, and J. Steinert, FCI normalized gain, scientific reasoning ability, thinking in physics, and gender effects, AIP Conf. Proc. **1413**, 23 (2012).
[116] V. P. Coletta and J. A. Phillips, Interpreting FCI scores: Normalized gain, preinstruction scores, and scientific reasoning ability, Am. J. Phys. **73**, 1172 (2005).
[117] T. L. McCaskey, M. H. Dancy, and A. Elby, Effects on assessment caused by splits between belief and understanding, AIP Conf. Proc. **720**, 37 (2004).
[118] T. L. McCaskey and A. Elby, Probing students' epistemologies using split tasks, AIP Conf. Proc. **790**, 57 (2005).
[119] 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).
[120] P. B. Kohl and H. V. Kuo, Introductory physics gender gaps: Pre-and post-studio transition, AIP Conf. Proc. **1179**, 173 (2009).
[121] 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).
[122] 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).
[123] L. Ding, Theoretical perspectives of quantitative physics education research, Phys. Rev. Phys. Educ. Res. **15**, 020101 (2019).
[124] 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).
[125] M. Meeuwisse, S. E. Severiens, and M. P. Born, Learning environment, interaction, sense of belonging and study success in ethnically diverse student groups, Res. High. Educ. **51**, 528 (2010).
[126] R. Masika and J. Jones, Building student belonging and engagement: Insights into higher education students' experiences of participating and learning together, Teach. High. Educ. **21**, 138 (2016).
[127] C. Goodenow, Classroom belonging among early adolescent students: Relationships to motivation and achievement, J. Early Adolesc. **13**, 21 (1993).
[128] A. W. Astin, *What Matters in College: Four Critical Years Revisited* (Jossey-Bass, San Francisco, CA, 1993).
[129] T. M. Freeman, L. H. Anderman, and J. M. Jensen, Sense of belonging in college freshmen at the classroom and campus levels, J. Exp. Educ. **75**, 203 (2007).

[130] 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).

[131] 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).

[132] 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).

[133] D. Hammer, Epistemological beliefs in introductory physics, Cognit. Instr. **12**, 151 (1994).

[134] Learning Activation Lab, Activation lab tools: Measures, and data collection instruments (2017), http://www.activationlab.org/tools/.

[135] W. K. Adams, K. K. Perkins, N. S. Podolefsky, M. Dubson, N. D. Finkelstein, and C. E. Wieman, New instrument for measuring student beliefs about physics and learning physics: The Colorado Learning Attitudes about Science Survey, Phys. Rev. ST Phys. Educ. Res. **2**, 010101 (2006).

[136] B. M. Zwickl, T. Hirokawa, N. Finkelstein, and H. J. Lewandowski, Epistemology and expectations survey about experimental physics: Development and initial results, Phys. Rev. ST Phys. Educ. Res. **10**, 010120 (2014).

[137] J. Schell and B. Lukoff, Peer instruction self-efficacy instrument, Developed at Harvard University, 2010 (unpublished).

[138] PERTS Academic Mindsets Assessment (2020), https://www.perts.net/orientation/ascend.

[139] 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, in *Proceedings of PER Conf. 2017, Cincinnati, OH*, 10.1119/perc.2017.pr.046.

[140] 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, in *Proceedings of PER Conf. 2017, Cincinnati, OH*, 10.1119/perc.2017.pr.064.

[141] T. J. Nokes-Malach, Z. Y. Kalender, E. Marshman, C. D. Schunn, and C. Singh, Prior preparation and motivational characteristics mediate relations between gender and learning outcomes in introductory physics, in *Proceedings of PER Conf. 2018, Washington, DC*, 10.1119/perc.2018.pr.Nokes-Malach.

[142] 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?, in *Proceedings of PER Conf. 2018, Washington, DC*, 10.1119/perc.2018.pr.Kalender.

[143] 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).

[144] 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).

[145] B. Thompson, *Exploratory and Confirmatory Factor Analysis* (American Psychological Association, Washington, 2004).

[146] K. Pearson and F. Galton, VII. Note on regression and inheritance in the case of two parents, Proc. R. Soc. London **58**, 240 (1895).

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

[148] R. Likert, A technique for the measurement of attitudes, Arch. Sci. Psychol. **22**, 55 (1932).

[149] A. Godwin, G. Potvin, and Z. Hazari, The development of critical engineering agency, identity, and the impact on engineering career choices, in *Proceedings of the 2013 ASEE Annual Conference & Exposition, Atlanta, Georgia* (2013), p. 23.1184. 1.

[150] E. A. Royse, E. Sutton, M. E. Peffer, and E. A. Holt, The Anatomy of Persistence: Remediation and Science Identity Perceptions in Undergraduate Anatomy and Physiology, Int. J. Teach. Learn. Higher Educ. **9**, 283 (2020).

[151] K. N. Hosbein and J. Barbera, Alignment of theoretically grounded constructs for the measurement of science and chemistry identity, Chem. Educ. Res. Pract. **21**, 371 (2020).

[152] X. Guo, X. Hao, W. Deng, X. Ji, S. Xiang, and W. Hu, The relationship between epistemological beliefs, reflective thinking, and science identity: A structural equation modeling analysis, Int. J. STEM Educ. **9**, 40 (2022).

[153] R. Dou, Z. Hazari, K. Dabney, G. Sonnert, and P. Sadler, Early informal STEM experiences and STEM identity: The importance of talking science, Sci. Educ. **103**, 623 (2019).

[154] R. Dou and H. Cian, Constructing STEM identity: An expanded structural model for STEM identity research, J. Res. Sci. Teach. **59**, 458 (2022).

[155] Z. Hazari, G. Potvin, R. M. Lock, F. Lung, G. Sonnert, and P. M. Sadler, Factors that affect the physical science career interest of female students: Testing five common hypotheses, Phys. Rev. ST Phys. Educ. Res. **9**, 020115 (2013).

[156] S. E. Embretson and S. P. Reise, *Item Response Theory for Psychologists* (Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US, 2000).

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

[158] R. P. Chalmers, Mirt: A multidimensional item response theory package for the R environment, J. Stat. Softw. **48**, 1 (2012).

[159] F. Samejima, Estimation of latent ability using a response pattern of graded scores, *Psychometrika Monograph* (Psychometric Society, Richmond, VA, 1969), p. 17.

[160] 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).

[161] 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).

[162] J. Pallant, SPSS Survival Manual: A Step by Step Guide to Data Analysis using IBM SPSS (Routledge, London, 2020).
[163] R Core Team, R: A Language and Environment for Statistical Computing (2013), (R Foundation) https://www.r-project.org/.
[164] A. J. Tomarken and N. G. Waller, Structural equation modeling: Strengths, limitations, and misconceptions, Annu. Rev. Clin. Psychol. **1**, 31 (2005).
[165] R. B. Kline, *Principles and Practice of Structural Equation Modeling* (Guilford Publications, New York, 2015).
[166] D. Kaplan, *Structural Equation Modeling: Foundations and Extensions* (Sage Publications, Thousand Oaks, CA, 2008), Vol. 10.
[167] D. Kaplan, Structural equation modeling, in *International Encyclopedia of the Social, and Behavioral Sciences*, edited by N. J. Smelser and P. B. Baltes (Pergamon, Oxford, 2001), p. 15215.
[168] Z. Awang, *SEM Made Simple: A Gentle Approach to Learning Structural Equation Modeling* (MPWS Rich Publication, 2015).
[169] 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).
[170] 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).
[171] 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, Psychological Methods **17**, 354 (2012).
[172] D. Hooper, J. Coughlan, and M. Mullen, Structural equation modeling: Guidelines for determining model fit, Electronic J. Bus. Res. Methods **6**, 53 (2007).
[173] 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, J. Anx. Dis. **27**, 735 (2013).
[174] C. A. Kronauge, *The Effects of Mixing Metrics and Distributions Simultaneously in Structural Equation Modeling: A Simulation Study* (University of Northern Colorado, Greeley, CO, 2012).
[175] 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).
[176] K. L. Tonso, Student engineers and engineer identity: Campus engineer identities as figured world, Cult. Stud. Sci. Educ. **1**, 273 (2006).
[177] J. E. Stets and P. J. Burke, A sociological approach to self and identity, *Handbook of Self and Identity* (Guilford Press, New York, 2003), p. 128.
[178] Z. Hazari, C. Cass, and C. Beattie, Obscuring power structures in the physics classroom: Linking teacher positioning, student engagement, and physics identity development, J. Res. Sci. Teach. **52**, 735 (2015).
[179] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences* (L. Erlbaum Associates, Hillsdale, NJ, 1988).
[180] 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).
[181] J. B. Grace and K. A. Bollen, Interpreting the results from multiple regression and structural equation models, Bulletin of the Ecological Society of America **84**, 283 (2005).
[182] R. M. Felder, G. N. Felder, and E. J. Dietz, A longitudinal study of engineering student performance and retention. V. Comparisons with traditionally-taught students, J. Engin. Educ. **87**, 469 (1998).
[183] R. Barnett, *The Ecological University: A Feasible Utopia* (Routledge, London, 2017).
[184] D. W. Johnson, R. T. Johnson, and K. A. Smith, Cooperative learning: Improving university instruction by basing practice on validated theory, J. Excellence Univ. Teach. **25**, 1 (2014).
[185] L. Bian, S.-J. Leslie, and A. Cimpian, Gender stereotypes about intellectual ability emerge early and influence children's interests, Science **355**, 389 (2017).
[186] S.-J. Leslie, A. Cimpian, M. Meyer, and E. Freeland, Expectations of brilliance underlie gender distributions across academic disciplines, Science **347**, 262 (2015).
[187] A. Deiglmayr, E. Stern, and R. Schubert, Beliefs in "brilliance" and belonging uncertainty in male and female STEM students, Front. Psychol. **10**, 1114 (2019).
[188] 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).
[189] 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).
[190] 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).
[191] 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).
[192] J. Wang and Z. Hazari, Promoting high school students' physics identity through explicit and implicit recognition, Phys. Rev. Phys. Educ. Res. **14**, 020111 (2018).
[193] A. S. Bryk and S. W. Raudenbush, *Hierarchical Linear Models: Applications and Data Analysis Methods* (Sage Publications, Thousand Oaks, CA 1992).