Examining the relation of high school preparation and college achievement to conceptual understanding

Dona Hewagallage, Elaine Christman, and John Stewart[*]

West Virginia University, Department of Physics and Astronomy, Morgantown, West Virginia 26506, USA

(Received 29 January 2022; accepted 13 June 2022; published 28 June 2022)

This study investigated factors influencing Force and Motion Conceptual Evaluation (FMCE) pretest and post-test scores for a sample ($N = 1116$ students) collected in the introductory calculus-based mechanics class at a large eastern land-grant university. Several academic and noncognitive factors were examined using correlation analysis and linear regression analysis to understand their relation to students' physics conceptual understanding. High school physics preparation was the most important factor in predicting FMCE pretest score. The kind of high school physics class (normal or Advanced Placement) and the student's academic performance in that class also greatly affected pretest scores. The optimal linear regression model explained 28% of the variance of pretest scores. Controlling for pretest score, ACT or SAT verbal and mathematics scores, students' grade expectation, and self-efficacy significantly predicted post-test score. The optimal linear regression model explained 54% of the variance of post-test scores. Pretest scores completely captured the effect of high school preparation on post-test scores; if pretest scores were included in a model predicting post-test scores, then high school physics preparation variables were not significant. Gender differences were observed on both the pretest and the post-test. These differences were not substantially mediated by either academic or noncognitive factors.

DOI: 10.1103/PhysRevPhysEducRes.18.010149

I. INTRODUCTION

The practice of comparing physics conceptual pretest and post-test scores to evaluate the development of conceptual understanding is prevalent in physics education research (PER) and is often used in physics classes in general. Most studies treat pretest and post-test scores as simple measures of physics conceptual knowledge. A growing body of research suggests this model is incomplete with studies demonstrating a relation between these scores and general academic preparation [1,2], demographic factors [3–5], and noncognitive factors (such as self-efficacy) [6].

Pretests often present students with an unusual testing situation where they are given an examination for very little course credit testing over material which has not been covered in class yet and for which they have not studied. The unusual nature of the testing situation may make noncognitive factors such as the student's self-beliefs or the student's personality more important than in more familiar testing situations. It is also possible that pretest scores incorrectly measure a student's actual prior preparation in physics because the student is not allowed to review for the pretest. These factors open the possibility that pretest scores do not fully capture prior preparation and that prior preparation has an additional effect on post-test scores not captured by pretest scores.

The practice of giving students a conceptual pretest prior to instruction and the same instrument as a post-test after instruction has been common in PER since the early days of the field. In 1985, Halloun and Hestenes used this methodology to show traditional instruction provided little additional conceptual understanding in college classes [7]. This observation led to the development of a catalog of student misconceptions about mechanics [8] which ultimately lead to the development of the Force Concept Inventory (FCI) [9] to measure conceptual understanding with an instrument that also presented students with commonly selected incorrect answers. Hake used pretest and post-test scores to show that the failure of traditional instruction to improve conceptual understanding was general [10]. The Hake study provided substantial impetus for the conversion to more inquiry driven modes of instruction. A 2014 synthesis by the National Academy of Sciences showed that these reformed modes of instruction improved performance on a broad collection of physics assessments including conceptual inventories [11]. The Hake study also popularized the use of the normalized gain which attempted to control for differing incoming levels of preparation by

[*] jcstewart1@mail.wvu.edu

dividing the absolute gain (post-test–pretest) by the available gain (100%–pretest).

The success of the FCI and the impact of the Hake study has led to the development of many conceptual instruments to probe student understanding on a variety of conceptual topics in physics. Among the more widely cited in research studies are the Force and Motion Conceptual Evaluation (FMCE) [12], the Conceptual Evaluation of Electricity and Magnetism (CSEM) [13], and the Brief Electricity and Magnetism Assessment (BEMA) [14]. Current versions of many assessments are available at PhysPort [15].

A. Research questions

This study was designed to explore the relation of high school preparation, college achievement, and noncognitive factors shown to be associated with college achievement (i.e., self-efficacy) and physics conceptual pretest scores. It also investigated how these factors as well as pretest scores influence post-test scores. Furthermore, this study seeks to provide a more thorough exploration of these factors than presented in prior works to extend the understanding of the incoming conceptual understanding of physics students.

This study seeks to answer the following research questions:

RQ1 What academic and noncognitive factors are most important in predicting FMCE pretest scores?

RQ2 What academic and noncognitive factors are most important in predicting FMCE post-test scores correcting for FMCE pretest scores?

RQ3 Do academic and noncognitive factors explain gender differences in FMCE pretest and post-test scores?

B. Pretest as a control

The use of pretest and post-test scores in PER is so common that any summary of prior research is necessarily incomplete. Some excellent review and synthesis articles can provide readers with an overview of the field and the role of research-based assessments in the field. McDermott and Redish provided an overview of early work in PER [16]. In 2014, Docktor and Mestre provided an extensive synthesis of research in PER [17]. In 2017, Madsen *et al.* provided an exhaustive overview of research-based assessment instruments in physics [18]. The instruments are often used to establish the efficacy of active teaching methods and other classroom interventions; Meltzer and Thornton provided an overview of different reformed instructional models and the research supporting the efficacy of those models [19]. The efficacy is often established by applying a pretest followed by a post-test.

Multiple large studies have shown either the efficacy of reform instruction across multiple institutions or the failure of traditional instruction to improve conceptual learning. Hake collected data from 62 physics classes at multiple institutions to show interactive instruction was superior to traditional instruction in promoting conceptual learning [10]. Von Korff *et al.* synthesized research using either the FCI or FMCE from 1995 to 2014 (a sample containing 50 000 students) to show that interactive instruction produced higher normalized gains than traditional instruction. Freeman *et al.* synthesized research from multiple scientific domains to show this result was general and not unique to physics classes [11]. A meta-analysis by Schroeder *et al.* demonstrated that reformed teaching methods are effective at promoting learning for students at many different points in their education [20]. Many studies have reported gender differences in conceptual pretest and post-test scores; Madsen *et al.* provided a summary of this research [3].

1. Gain scores

Pretest–post-test designs are used in studies in many different fields to understand the effectiveness of a treatment. This design has been analyzed in different ways to characterize the overall change [21]. Within PER, the normalized gain, the ratio of actual gain to the maximum possible gain, is often reported. This statistic was popularized in an influential study by Hake comparing instructional methods [10]. Nissen *et al.* showed the normalized gain was biased in favor of populations with higher pretest scores and suggested an alternate gain score using Cohen's d [22]. Either the actual gain, the normalized gain, or Cohen's d depend on the pretest score, the post-test score, and the relation of pretest score to post-test score. As such, all may be influenced by factors related to any of these quantities.

2. Demographics and conceptual inventory scores

Many studies have reported and explored differences between the conceptual inventory pretest or post-test scores of members of demographic subgroups and nonmembers of those groups including underrepresented minority students (URM), first-generation college students (FGCS), women, and rural students. Most of these studies have examined differences by gender, but more recent studies have investigated other groups.

Salehi *et al.* examined performance differences in introductory physics between several demographic groups 1]]. Differences in final exam scores between demographic groups were fully explained by differences in SAT scores and pretest scores. The study investigated three samples; two used the FMCE as the pretest, one the FCI as a pretest. Stewart *et al.* partially replicated this work examining performance differences in FMCE post-test scores and course grades [2]. General high school preparation measured by ACT and SAT scores and prior preparation in physics strongly mediated demographic performance differences for FGCS and URM students on both post-test scores and grades. No difference in course grades between men and women existed, so no mediation was possible. Gender differences in post-test scores were weakly mediated by ACT and SAT score and pretest scores with much of the initial gender difference unexplained by these factors. Henderson *et al.* examined the amount of the gender gap that was explained by instrumental fairness, ACT and SAT scores, and pretest scores in five large samples including two FMCE samples [4] finding that different factors affect post-test scores in the five samples by different amounts, but in all samples a large part of the post-test gender differences were unexplained by these factors. Other studies have also found differences between rural and nonrural students on the FMCE pretest and post-test [5]. Pretest scores on the FCI, the FMCE, and the CSEM also correlate with postinstruction achievement measures (post-test score, test average, and course grades) differently for members of different demographic groups [23]. As such, general high school measures of achievement, ACT and SAT scores, and measures of prior physics knowledge explain some variation in a variety of physics achievement measures, but the variation explained is not consistent for different groups and much of the variation in the conceptual post-test performance of women is unexplained.

C. Factors influencing pretest scores

Many studies have investigated factors outside the college physics classroom influencing pretest scores, post-test scores, and normalized gains including demographic factors, general high school academic factors, and specific high school instruction in physics. Most of these studies have focused on class grades, test averages, post-test scores, and normalized gains; however, it seems quite likely that student factors that existed prior to taking the physics class might also influence pretest scores. Support for this can be found in recent studies presenting path models including pretest scores, standardized test scores (ACT or SAT), and class outcome variables (grades, final exam scores, or post-test scores) showing the ACT and SAT scores have a significant effect on pretest scores as well as an effect on class outcomes controlling for pretest scores [1,2].

Early work in PER predating the FCI investigated the effect of many cognitive factors on course grades or test averages including formal operational reasoning [24,25], mathematics pretest scores [25,26], and logical reasoning [26]. Meltzer showed the normalized gain on an electricity conceptual inventory was correlated with mathematics pretest scores and ACT or SAT mathematics percentile scores [27]. Coletta and Phillips found a positive correlation between Lawson's Classroom Test of Scientific Reasoning and FCI normalized gains [28]. Coletta *et al.* demonstrated a strong positive correlation between composite SAT scores and normalized gains on the FCI in both college-level and high-school-level students [29].

In an unpublished work but highly cited work, Hake showed that having high school physics affected college physics normalized gains on the FCI, but the effect was a small effect ($d = 0.19$) [30]. According to Hart and Cottle, math proficiency and high school physics background are vital for college achievement [31]. Hazari *et al.* investigated the relation of high school mathematics and sciences grades, taking Advanced Placement (AP) calculus, instructional format, and some noncognitive factors involving family support and found that many of these factors significantly predicted physics grades in college [32] controlling for demographic characteristics. Kost *et al.* explored the effect on post-test scores controlling for pretest scores and gender of many factors including mathematics preparation measured both with standardized test scores and a university-applied placement test and students' attitudes about science finding both sets of variables as significant predictors of post-test scores [33]. They also report a 7% difference in FMCE post-test scores between students who had high school physics and students who did not; the affect of high school physics was larger for women, a 14% difference.

D. Studies of the properties of the FMCE

Many studies have examined the item properties of the FMCE including their factor structure [34,35], their network structure [36], and their psychometric properties [34,37,38]. Psychometric properties investigated include reliability, problematic item functioning, and item bias. More qualitative analyses have examined the instrument through the lens of the resource framework [39].

Ramlo examined the factor structure and reliability of the FMCE using a sample of 146 students [34]. The instrument was reliable with Cronbach's alpha of 0.742 for the pretest and 0.907 for the post-test. Ramlo found the pretest factors extracted mixed items testing different concepts and thus concluded that the pretest factor structure was undefined. The post-test factor structure contained three factors. Yang *et al.* examined the post-test factor structure using multidimensional item response theory (MIRT) and found 5 factors as optimal [35]. These factors also contained loadings mixing different topics in mechanics.

Henderson *et al.* examined the item characteristics of the FMCE using classical test theory (CTT) and differential item functioning (DIF) theory disaggregating the sample by gender [37]. Many FMCE items had difficulty or discrimination within the range of problematic item functioning using CTT [40] on the pretest; fewer items were problematic on the post-test. Unlike the FCI, which contained many items unfair to women (and a few unfair to men) identified using DIF [41], the FMCE contained only one unfair item identified in both samples and this item was unfair to men (one sample contained a single item unfair to women).

E. Factors affecting college achievement

Pretest and post-test scores measure a student's knowledge of physics. The research into the factors affecting pretest score summarized above also show they measure other academic factors such as general high school preparation. As such, they may be related to factors identified as important in college achievement in general.

1. General academic factors

A substantial strand of education research seeks to understand the factors that influence academic achievement at the college level. Much work has been focused on SAT and ACT scores as predictors of college achievement. Composite scores on the SAT and ACT are highly correlated with each other [42] and with measures of general cognitive ability [43,44]. The College Board touts the SAT's validity as a predictor of freshman-year GPA [45], while ACT has developed benchmarks for scores indicating a 50% chance of earning a B or higher in introductory college courses [46].

Although high school grades offer a less standardized measure of academic performance due to differing grading practices in different classrooms and schools [47], they are consistently stronger predictors of freshman-year GPA [42,48], cumulative college GPA [49], and college completion [49–51] than ACT and SAT scores. Galla *et al.* found that self-regulation explained far more of the variance in students' high school GPA than did cognitive ability and that this in turn explained the greater incremental predictive validity of high school grades over ACT and SAT scores for college completion [51].

2. Noncognitive factors

Many research studies have explored the influence noncognitive factors on college achievement including personality traits, motivational factors, and psychosocial contextual influence [52,53].

Self-efficacy, "people's beliefs about their capabilities to produce designated levels of performance that exercise influence over events that affect their lives" as defined by Bandura [54], has been shown to affect student's performance and achievement in science classes [52,55,56]. Many studies have found that male students have higher self-efficacy than female students in science, technology, engineering, and mathematics (STEM) classes [57–60]. Besterfield-Sacre *et al.* showed that these differences exist at the beginning of college using a study at 17 institutions [61]. Dou *et al.* reported that, regardless of gender, students on average had lower self-efficacy at the end of the semester compared to the beginning of the semester [62]. In physics, a study exploring impact on Modeling Instruction on self-efficacy, reported that traditional lecture classrooms negatively impact self-efficacy [63]. Cwik and Singh reported a decrease in the self-efficacy gender difference from the beginning to the end of the course and that it was not due to the difference of performance between men and women [64].

Hagerty *et al.* defined sense of belonging as "the experience of personal involvement in a system or environment so that persons feel themselves to be an integral part of that system or environment." Sense of belonging has been an important construct in STEM education research and has been shown to be related to students' achievement [65]. The relation of belonging to achievement has been investigated in many STEM domains. Sense of belonging in mathematics classes was predictive of mathematics achievement after controlling for other constructs affecting performance [66]. In physics, sense of belonging has been shown to relate to physics achievement (course grades) and participation; the relation was the same for men and women. Stereotype endorsement affected the sense of belonging in physics for women with ACT and SAT mathematics scores positively related to belonging [67]. Lewis *et al.* showed that exposure to stereotypical cues decreased women's sense of belonging while the presence of other female peers and role models impacted belonging positively [68]. Belonging interventions increased the sense of belonging of women in engineering majors where they are substantially underrepresented [69].

In this and many works, personality was characterized using the five-factor model with facets: agreeableness, conscientiousness, extraversion, neuroticism, and openness [70–72]. The model is usually measured with an instrument using Likert scale items. Personality has been shown to have a direct influence on academic performance and achievement [52,53]. Stewart *et al.* reported that students' self-efficacy and personality were related to their college achievement [73]. Each facet measures a distinct characteristic of personality; as such, their interactions with academic performance also differ. Agreeableness, an individual's tendency to be cooperative and compassionate, has a positive correlation with academic performance. Similarly conscientiousness, how organized, focused, and careful an individual is, also positively correlates with achievement. Openness, one's willingness to embrace new ideas and experiences, correlates positively with academic performance. Unlike the previous facets, extraversion, one's inclination for social interactions and attention, negatively correlates with academic performance. Neuroticism, how anxious one feels, also negatively correlates with academic performance [52].

Beyond the noncognitive factors examined in the present study, many studies have investigated other factors that may affect performance in the STEM classroom and how these factors may explain demographic differences in performance. Other extensively explored noncognitive factors include mathematics anxiety [74,75], science anxiety [76–78], stereotype threat [79], and attitudes toward science, see Table I in Ref. [3]. Theoretically, the noncognitive factors explored in the present study should possibly be related to some of these additional factors but additional research is required to establish and understand the relation.

The purpose of this study is threefold with the overall goal of deepening the understanding of the factors which influence both conceptual pretest and post-test scores. First, conceptual pretest scores, while broadly used, have received far less study than post-test scores. Theoretically, pretest scores could be influenced by a broad set of both cognitive and noncognitive variables. Because pretest scores are routinely used as control variables in PER studies, it is crucial that their relations to other constructs are well understood. This study collects a rich dataset containing both noncognitive variables theoretically related to exam performance and a more detailed set of high school preparation variables than have previously been used in PER studies. Second, this study seeks to determine how completely pretest scores measure the effect of these variables on post-test scores. If pretest scores incompletely capture the effects of high school preparation on post-test scores, their use as control variables is inaccurate. Third, using the broad set of variables available to the study, this work seeks to shed light on one of the most studied outstanding questions in PER. What is the origin of the gender differences observed in conceptual inventory scores?

II. METHODS

A. The FMCE

The FMCE [12] measures conceptual understanding of Newtonian mechanics. The test consists of 43 multiple choice items (excluding the energy items). After its introduction, Thornton *et al.* [80] introduced a modified scoring method that produced a total score of 33 by eliminating some items and scoring some items as groups; this method is used in the current study.

B. Sample

This study was performed from Fall 2017 to Fall 2019 at a large land-grant university in the eastern United States. The university's general undergraduate population was 80% White, 6% international, 4% Hispanic, 4% African American, 4% students reporting two or more races, 2% Asian, and other groups each with 1% or less [81].

The study was performed in the calculus-based introductory mechanics course taken by scientists and engineers. The class was presented with three 50-min lecture sessions and one 3-h required laboratory session each week. The class was overseen by a single lead instructor with a strong knowledge of research-based instruction and a commitment to student engagement. This instructor managed lab and homework content. The class offered multiple lecture sections each semester either taught by the lead instructor or by another instructor in partnership with the lead instructor. All lecture sessions implemented the Peer Instruction [82] pedagogy using clickers. The laboratory session featured a mix of white-boarding activities, hands-on inquiry activities, traditional experiments, and group problem solving.

Student demographic and college performance measures were accessed from institutional records. Noncognitive factors were measured using a survey instrument given the first week of the semester. Student high school science and mathematics course information was collected using a survey instrument given the second week of the semester.

In the period studied, 3777 students enrolled in the class studied. Removing students without basic high school information (GPA, ACT, or SAT scores), students not enrolled as first-time freshman, and students without college level academic information such as college GPA left 3063 students. Removing students without FMCE pretest or post-test scores left 2279 students. Students were also removed who did not take both of the survey instruments leaving an overall sample size for this study of $N = 1116$.

This study was performed using the same general student population examined in some of the PER studies examined above [2,4,5,23,37,73].

C. Instruments

Noncognitive and high school course taking were accessed using two surveys given early in the semester. Some survey items were constructed for this study and some were taken from published work.

This study collected four noncognitive measures which could be related to pretest performance: personality, self-efficacy, sense of belonging, and the student's self-reported expected grade (grade expectation) in the class studied. Self-efficacy, a student's belief that he or she will be successful in the class, could influence the student's performance by modifying the effort invested in the pretest or by modifying patterns of answering pretest questions, possibly causing low self-efficacy students to doubt their answers (and possibly change those answers). Self-efficacy has long been reliably associated with academic performance [52]. For the class studied, the pretest is given the first week of the semester before any physics content has been covered. The student is in their first college physics class, in an unfamiliar setting, generally with many students he or she does not know. Whether the student feels they belong in this setting could possibly influence how he or she performs on a lightly incentivized exam over material not yet covered in the class by modifying the effort directed toward the exam or by encouraging the student to complete the exam quickly so he or she could leave the setting. Personality, measured by the five-factor model, has two facets of particular interest in the pretest setting: neuroticism and conscientiousness. Conscientiousness, the tendency to carefully complete tasks, could adjust the care taken with the pretest. Conscientiousness has also been reliably shown to correlate with academic achievement [52]. Neuroticism, the tendency to feel stress or other strong emotions, could influence a students reaction to the unfamiliar testing situation. Finally, the student's grade expectation which is related to both self-efficacy and academic motivation, could modify how effectively the incentive given for the pretest causes careful work.

1. Personality

Personality was measured using the big five inventory (BFI) which uses five facets to characterize personality: agreeableness, conscientiousness, extraversion, neuroticism, and openness [70–72]. It contains 44 survey questions with each measured on a five-point Likert scale. The BFI has been extensively used in a broad variety of research [83].

2. Self-efficacy

Self-efficacy was measured using the self-efficacy for learning and performance subscale from the motivated strategies for learning questionnaire (MSLQ) [84]. The subscale has strong validity [84] and is widely used [85]. This subscale asks the student to rate how much they agree with statements accessing self-efficacy on a 5-point Likert scale. For example, "I'm confident I can do an excellent job on the assignments and tests in this course." These statements were specialized by replacing "course" with "physics class." Word substitution to specialize the MSLQ to specific domains has been used in previous research studies [86].

3. Belonging

A student's sense of belonging in his or her physics class was accessed using three items adapted from Good, Rattan, and Dweck's "Math Sense of Belonging" instrument [66]. For example, students were asked how much they agree with the statement "I feel I fit in when I am in physics classes and with students in my physics classes." One's sense of belonging in a class could affect performance on an examination either by reducing or increasing anxiety or changing one's belief that they could succeed on an examination.

Both the sense of belonging and the self-efficacy subscales were modified from their original published form. Both subscales were re-validated before the study began first by conducting interviews with students to confirm the modified items were being interpreted correctly, then by applying open-ended versions of the items and examining responses. The internal reliability of the final version of each subscale was characterized by Cronbach's alpha showing each subscale was highly reliable ($\alpha > 0.9$).

4. Grade expectation

Students were also asked to predict the grade they would receive in the class using the question "What grade do you expect to get in your physics class?" This was converted to a three-level variable: "A," "B," and "C, D, F, or W."

Personality, self-efficacy, belonging, and grade expectation were collected with a survey instrument given in the first full week of classes. Students received a small amount of course credit upon the completion of the survey.

5. High school preparation

High school physics and mathematics programs are highly variable in how well they prepare students for college. Universities often collect incomplete information about high school course taking (or store such information in digitally inaccessible forms, such as images). To collect more complete information, students were given a survey instrument that asked about high school science and mathematics preparation in detail.

Information on AP and transfer classes was available for the institution studied. This was only available for AP or transfer (dual enrollment) classes which received college credit (a minimum AP score or a passing transfer grade). All students retained in the sample were enrolled as "first-time freshmen" and, therefore, transfer classes were taken in high school. To capture AP classes taken where the AP test was not passed, the students were also asked to report the AP mathematics and physics classes taken and to report their score on the AP test.

Students were asked about the first and second high school science classes taken in each of three domains: physics, chemistry, and biology. They were also asked to classify the level of each class as "regular," "honors," "AP," "dual enrollment," and "other advanced" and asked to report the grade they received in each class. This generated a very complex set of data with many of the categories implied by the many levels in the data containing few students. Preliminary analysis first fit the raw survey data predicting pretest score, then formed combinations of variables to yield a more parsimonious set of variables with similar predictive power where all levels of each variable contained enough students for statistical reliability. This resulted in a seven-level categorical variable HS Physics which combined broad divisions of the type of the last high school physics class taken with the grade in the class. These two measures were combined because a student who has a grade in high school physics has taken high school physics and we wanted to isolate the overall effect of taking a high school physics class. Grades were divided into two levels, "A" and "B, C, or D"; types of physics class where divided into "high school physics not taken," "high school physics not AP," "high school physics AP—test not passed (no college credit)," and "high school physics AP—test passed." Multiple AP high school physics classes are offered; students with credit for the calculus-based class are not required to take the class studied. As such, students with AP physics credit had taken the algebra-based AP physics class.

All students reporting HSGPA taking a college physics class had taken some mathematics in high school. Students were asked to report the most advanced high school class taken and the grade in that class. A similar procedure of analysis yielded two variables: a dichotomous "high school last math grade A" variable and a 4-level categorical variable capturing the type of most advanced high school mathematics class: "high school math not calculus," "high school math calculus—not AP," "high school math calculus—AP (test not passed)," and "high school math calculus—AP (test passed)." Note, regular, honors, and dual enrollment (concurrent credit) high school classes were combined into the high school physics (calculus) not AP category because disaggregating these levels did not improve predictive power. At the institution studied, a score of 4 on either the AP calculus or physics class was required to earn university credit (to pass the AP test).

For both mathematics and physics, passing the AP test was accessed from university records, not from the self-reported survey responses.

The variables described above focus on AP class taking. Students also receive college credit by taking college level classes while in high school; these class are called transfer classes. The number and type of transfer classes were very weakly predictive of pretest scores and were, therefore, not included in our final high school physics variable encoding.

D. Variables

Table I shows all variables used in this study. A short name is provided for each variable as well as a more complete description. The variables are divided into two types continuous (C) or dichotomous (D). Continuous variables are normalized by subtracting the mean and dividing by the standard deviation when used in linear regression analysis. By normalizing the continuous variables, which are all measured on different scales, one converts the scale of measurement of all variables to standard deviation units. This allows comparison of the importance of the different variables.

ACT and SAT scores were accessed from institutional records. Each was converted to a percentile score using tables published by the testing companies. When both were available, the two scores were averaged. The ACT English subscore was used as the ACT verbal score.

All calculations were performed with the "R" software system [87].

III. RESULTS

A. Descriptive statistics

Table II presents descriptive statistics for all variables. For dichotomous variables, the percentage of the students in the higher level of the variable (the student is in the state represented by the variable) is shown. For continuous variables, the mean and standard deviation of the variable is presented. The correlation of each variable with FMCE pretest score r and the significance of this correlation is presented. If the variable is continuous then the Pearson correlation is used; if dichotomous the point-biserial correlation. Variables are separated into groups that are called "panels" in this work. Some dichotomous variables are independent such as whether the student is repeating the physics class; some are not. For groups of interdependent dichotomous variables such as the variables in the high school (HS) physics panel, a base level of the variable is selected (indicated by "BL" in the Table II). Analyses calculate changes against this variable. For a dichotomous variable in a panel, the correlation for a nonbase variable is deceptive if calculated naively. For example, the high level of the variable "High school physics class not AP—A" represents students who took high school physics, but not as an AP class, and earned an A in the class. The low level of this variable represents all other students including students who did not take high school physics as well as students who took AP physics and passed the AP test. For a fair comparison of the importance of being in the high school physics class not AP—A group, students in this group are compared to the base level (students without high school physics) by subsetting the data to only include students in these two groups. Other nonbase variables in panels were handled similarly. The table also presents the R^2 values for a model regressing all variables in the panel on pretest score as well as the significance of the model. For panels with a single variable, $R^2_{\mathrm{panel}} = r^2$.

For correlation coefficients, Cohen's effect size criteria are $r = 0.1$ as a small effect, $r = 0.3$ as a medium effect, and $r = 0.5$ as a large effect [88]. Only a few variables have correlations with a medium to large effect; a number of variables in the HS physics panel meet this criteria. The effect of taking an AP physics class and earning an A in that class is substantial whether or not the AP test is passed (the effect is larger if the test is passed). The only other variable meeting this criteria is taking AP calculus and passing the AP test. A number of variables fall in the range $0.2 < r < 0.3$ (small to medium effects) including college math readiness, ACT or SAT math and verbal scores, self-efficacy, and reporting expecting to earn an A in the physics class. Not taking high school physics was negatively correlated with pretest scores ($r = -0.22$). Many variables exceeded the small effect size threshold including gender. General college success measured by college GPA was less correlated with pretest scores than the variables above implying that the pretest is measuring elements of preparation prior to entering college as opposed to general academic success in college. As such, the FMCE pretest seems to measure first high school preparation in physics (and the details of that preparation), then general high school preparation.

Each dichotomous variable divides the sample into two groups. Table III presents the number of students in each group, the mean and standard deviation, as well as the p value for a t test comparing the pretest scores of the two

TABLE I. List of variables. Type indicates whether the variable is continuous (C) or dichotomous (D). The base level variable for each dummy-coded multilevel variable is indicated by bold face.

Panel	Abbreviation	Type	Description
	Pretest	C	FMCE pretest percentage.
	Post-test	C	FMCE post-test percentage.
Repeat	Repeat	D	Is the student repeating the class?
College	Complete	C	Percentage of college classes attempted that are completed before class.
	CGPA	C	College grade point average before class.
	STEMCls	C	STEM classes completed before class.
	Credit	C	Credit hours completed before class.
	Enroll	C	Current hours enrolled in semester of physics class.
Math ready	MathReady	D	Was the student's first college mathematics class Calculus 1 or higher?
HS general	ACTM	C	ACT or SAT mathematics percentile score.
	ACTV	C	ACT English or SAT verbal percentile score.
	HSGPA	C	High school grade point average.
AP general	AP.NMP	D	Does student have AP credit excluding math and physics credit?
	AP.C.NMP	C	Number of non-math or non-physics classes with AP college credit.
Transfer	TR.NMP	D	Does the student have transfer credit excluding math and physics credit?
	TR.C.NMP	C	How many non-math and non-physics transfer classes?
	TR.Phys	D	Does the student have transfer credit for physics?
	TR.Math	D	Does the student have transfer credit for math?
HS physics	**HSP.NTake**	D	High school physics not taken.
	HSP.NAP.NA	D	High school physics class not AP—grade B, C, D.
	HSP.NAP.A	D	High school physics class not AP—grade A.
	HSP.APNP.NA	D	High school physics AP (test not passed)—grade B, C, D.
	HSP.APNP.A	D	High school physics AP (test not passed)—grade A.
	HSP.APP.NA	D	High school physics AP (test passed)—grade B, C, D.
	HSP.APP.A	D	High school physics AP (test passed)—grade A.
HS math	HSM.A	D	Was the grade in the student's most advanced high school math class an A?
	HSM.NCal	D	Was most advanced high school math class below calculus?
	HSM.NAP	D	Was most advanced high school math class calculus?
	HSM.APNP	D	Was most advanced high school math class AP calculus (test not passed)?
	HSM.APP	D	Was most advanced high school math class AP calculus (test passed)?
Belonging	Belong	C	Sense of belonging in physics class.
Self-efficacy	SelfEff	C	Self-efficacy towards physics class.
Grade expectation	GrdExA	D	Does the student expect to earn an A in physics?
	GrdExB	D	Does the student expect to earn a B in physics?
	GrdExC	D	Does the student expect to earn a C, D, F, or W in physics?
Personality	Agr	C	Personality facet—Agreeableness
	Cns	C	Personality facet—Conscientiousness
	Nrt	C	Personality facet—Neuroticism
	Ext	C	Personality facet—Extraversion
	Opn	C	Personality facet—Openness
Demographics	Gender	D	Does the student identify as female?
	FirstGen	D	Is the student a first-generation college student?
	URM	D	Does the student identify as URM?

groups. The effect size of the difference between pretest scores for the two levels of the variable is characterized by Cohen's d. Cohen's criteria for d are that 0.2 is a small effect, 0.5 is a medium effect, and 0.8 is large effect. While both effect sizes, the effect size criteria for r discussed earlier are effect sizes for the degree of association between two variables while Cohen's d measures the effect size of the difference between two groups.

TABLE II. Descriptive statistics. The base level of a set of dummy-coded variables is given by BL and indicated in bold face. For dichotomous variables, the percentage of students in the high level of the variable is reported. For continuous variables, the mean (M) and standard deviation (SD) is presented. For all variables, the correlation r with pretest score and the probability that the correlation or a larger correlation occurred by chance p are reported.

Panel	Variable	BL	%	$M \pm$ SD	r	p	R^2_{panel}	p_{panel}
	FMCE Pretest %			23.31 ± 18.3	1.00	0.000		
	FMCE Post-test %			46.61 ± 27.8	0.66	0.000		
Repeat	Is repeating physics class?		4.9		0.01	0.664	0.000	0.664
College	College course completion %			94.01 ± 11.1	0.11	0.000	0.055	0.000
	College GPA			3.35 ± 0.48	0.17	0.000		
	College STEM classes taken			3.78 ± 0.91	−0.10	0.001		
	College credit earned			27.11 ± 15.9	−0.16	0.000		
	College hours currently enrolled			16.6 ± 1.67	0.04	0.150		
Math ready	Entered college math in calculus		64.0		0.21	0.000	0.046	0.000
HS general	ACT or SAT mathematics %			81.21 ± 13.9	0.27	0.000	0.093	0.000
	ACT or SAT verbal %			75.32 ± 17.9	0.23	0.000		
	High school GPA			3.9 ± 0.44	0.05	0.073		
AP general	Has AP credit (not math or physics)		37.4		0.08	0.010	0.013	0.001
	Number AP classes (not math or physics)			4.15 ± 3.2	0.12	0.012		
Transfer	Has transfer credit (not math or physics)		35.5		−0.02	0.438	0.004	0.342
	Number transfer credits (not math or physics)			4.25 ± 4.2	−0.05	0.353		
	Has transfer credit physics		1.9		−0.01	0.840		
	Has transfer credit calculus		9.7		−0.06	0.050		
HS physics	**High school physics not taken**	×	22.0		−0.22	0.000	0.175	0.000
	High school physics class not AP—B, C, D		17.6		0.13	0.008		
	High school physics class not AP—A		31.9		0.21	0.000		
	High school physics AP (test not passed)—B, C, D		11.4		0.35	0.000		
	High school physics AP (test not passed)—A		13.4		0.46	0.000		
	High school AP physics test passed—B, C, D		0.9		0.35	0.000		
	High school AP physics test passed—A		2.9		0.65	0.000		
HS math	High school last math grade A		58.4		0.08	0.007	0.049	0.000
	High school last math not calculus	×	28.9		−0.15	0.000		
	High school last math calculus (not AP)		19.2		0.08	0.064		
	High school last math AP calculus (test not passed)		41.0		0.16	0.000		
	High school last math AP calculus (test passed)		10.8		0.32	0.000		
Belonging	Sense of belonging in physics			4.08 ± 0.69	0.12	0.000	0.015	0.000
Self-efficacy	Self-efficacy toward physics			4.06 ± 0.71	0.20	0.000	0.042	0.000
Grade expectation	Physics grade expectation A		41.3		0.24	0.000	0.044	0.000
	Physics grade expectation B		41.0		0.12	0.002		
	Physics grade expectation C, D, F, W	×	17.7		−0.15	0.000		
Personality	Agreeableness			3.84 ± 0.57	−0.06	0.032	0.031	0.000
	Conscientiousness			3.77 ± 0.55	−0.04	0.151		
	Neuroticism			2.79 ± 0.76	−0.02	0.606		
	Extraversion			3.19 ± 0.74	−0.12	0.000		
	Openness			3.65 ± 0.53	0.07	0.024		
Demographics	Gender (Female = 1)		29.1		−0.13	0.000	0.019	0.000
	First-Generation (First-gen = 1)		15.9		−0.05	0.077		
	URM (URM = 1)		7.0		−0.01	0.816		

TABLE III. Comparison of dichotomous variables. The levels of the variables are 0 or 1 and are indicated by subscripts. N_i represents the number of students in each level. The mean (M_i) and standard deviation for each level of the variable on the FMCE pretest is also presented. A t test was performed testing the difference between the levels. The significance of the t test is measured by the probability p and the effect size of the difference by Cohen's d.

Variable	N_0	N_1	$M_0 \pm$ SD	$M_1 \pm$ SD	p	d
Is repeating physics class?	1061	55	23.3 ± 18	24.4 ± 17	0.638	0.06
Entered college math in calculus	402	714	18.1 ± 14	26.3 ± 20	0.000	0.46
Has AP credit (not math or physics)	699	417	22.2 ± 17	25.1 ± 20	0.012	0.16
Has transfer credit (not math or physics)	720	396	23.6 ± 18	22.7 ± 18	0.436	0.05
Has transfer credit physics	1095	21	23.3 ± 18	22.5 ± 13	0.773	0.04
Has transfer credit calculus	1008	108	23.7 ± 19	20.0 ± 14	0.017	0.20
High school physics not taken	870	246	25.4 ± 19	15.9 ± 11	0.000	0.53
High school physics class not AP—B, C, D	920	196	15.9 ± 11	19.1 ± 15	0.010	0.25
High school physics class not AP—A	760	356	15.9 ± 11	22.2 ± 16	0.000	0.44
High school physics AP (test not passed)—B, C, D	989	127	15.9 ± 11	26.3 ± 17	0.000	0.79
High school physics AP (test not passed)—A	967	149	15.9 ± 11	33.8 ± 24	0.000	1.06
High school AP physics test passed—B, C, D	1106	10	15.9 ± 11	37.9 ± 22	0.011	1.91
High school AP physics test passed—A	1084	32	15.9 ± 11	53.1 ± 27	0.000	2.70
High school last math grade A	464	652	21.6 ± 16	24.6 ± 20	0.005	0.16
High school last math not calculus	793	323	25.1 ± 19	19.0 ± 14	0.000	0.33
High school last math calculus (not AP)	902	214	19.0 ± 14	21.5 ± 17	0.072	0.16
High school last math AP calculus (test not passed)	658	458	19.0 ± 14	24.8 ± 19	0.000	0.34
High school last math AP calculus (test passed)	995	121	19.0 ± 14	32.3 ± 24	0.000	0.77
Physics grade expectation A	655	461	17.5 ± 12	27.5 ± 21	0.000	0.53
Physics grade expectation B	658	458	17.5 ± 12	21.5 ± 17	0.000	0.26
Physics grade expectation C, D, F, W	919	197	24.5 ± 19	17.5 ± 12	0.000	0.39
Gender (Female = 1)	791	325	24.8 ± 19	19.7 ± 15	0.000	0.28
First-generation (First-gen = 1)	939	177	23.7 ± 19	21.1 ± 14	0.031	0.15
URM (URM = 1)	1038	78	23.3 ± 18	22.8 ± 19	0.820	0.03

Table III provides support for the observations made about Table II. Whether the student took high school physics represented a medium effect ($d = 0.53$) which is approximately commensurate with the effect of being calculus ready upon entering college ($d = 0.46$) and expecting to earn an A in the physics class ($d = 0.53$). Therefore, while taking high school physics is very important to pretest score, it is not uniformly the most important effect. Taking AP high school calculus and passing the AP test was a larger effect ($d = 0.77$).

The kind of high school physics taken has a dramatic effect on pretest scores, with taking AP physics increasing pretest scores from $d = 0.79$ to an extraordinary $d = 2.70$ for students who passed the AP test and report earning an A in the AP class. Comparisons of the mean percent score for different modes of taking high school physics and different grade outcomes also show exceptional differences with students who do not take high schools physics scoring 16% on the pretest and students who passed the AP test and earned an A scoring 53% on the pretest. As in Table II, for variables in a panel, the mean of the low level is the mean of students in the base level of the panel; for high school physics, the base level is students who do not take high school physics.

In summary, Table II shows that variables related to the details of a students high school experience, the type of high school physics taken and the student's performance in that class, as well whether the student took AP calculus and passed the AP test were the most correlated with pretest scores. As such, studies attempting to relate high school experiences to physics conceptual knowledge when entering a physics class should collect very detailed measures of that experience. More general measures of high school preparation, ACT and SAT scores, were also strongly correlated with pretest scores as were measures of self-efficacy, but not as strongly as physics and mathematics high school preparation. Table III supports these general conclusions with the strongest differences in pretest scores related to high school physics and mathematics differences.

B. Correlation analysis

Figure 1 shows a visualization of the correlation matrix for variables in Table I that are not part of the same panel. The visualization uses green (solid) lines for positive correlations and red (dashed) lines for negative correlations. Thicker lines represents a larger absolute value of the correlation. The visualization is rendered with the "qgraph" package in R that uses the force-direct graph visualization [89]. This representation is largely for visual effect, but it does allow the identification of groups of variables that are strongly intercorrelated. To produce the visualization, Hooke's lawlike attractive forces are introduced between

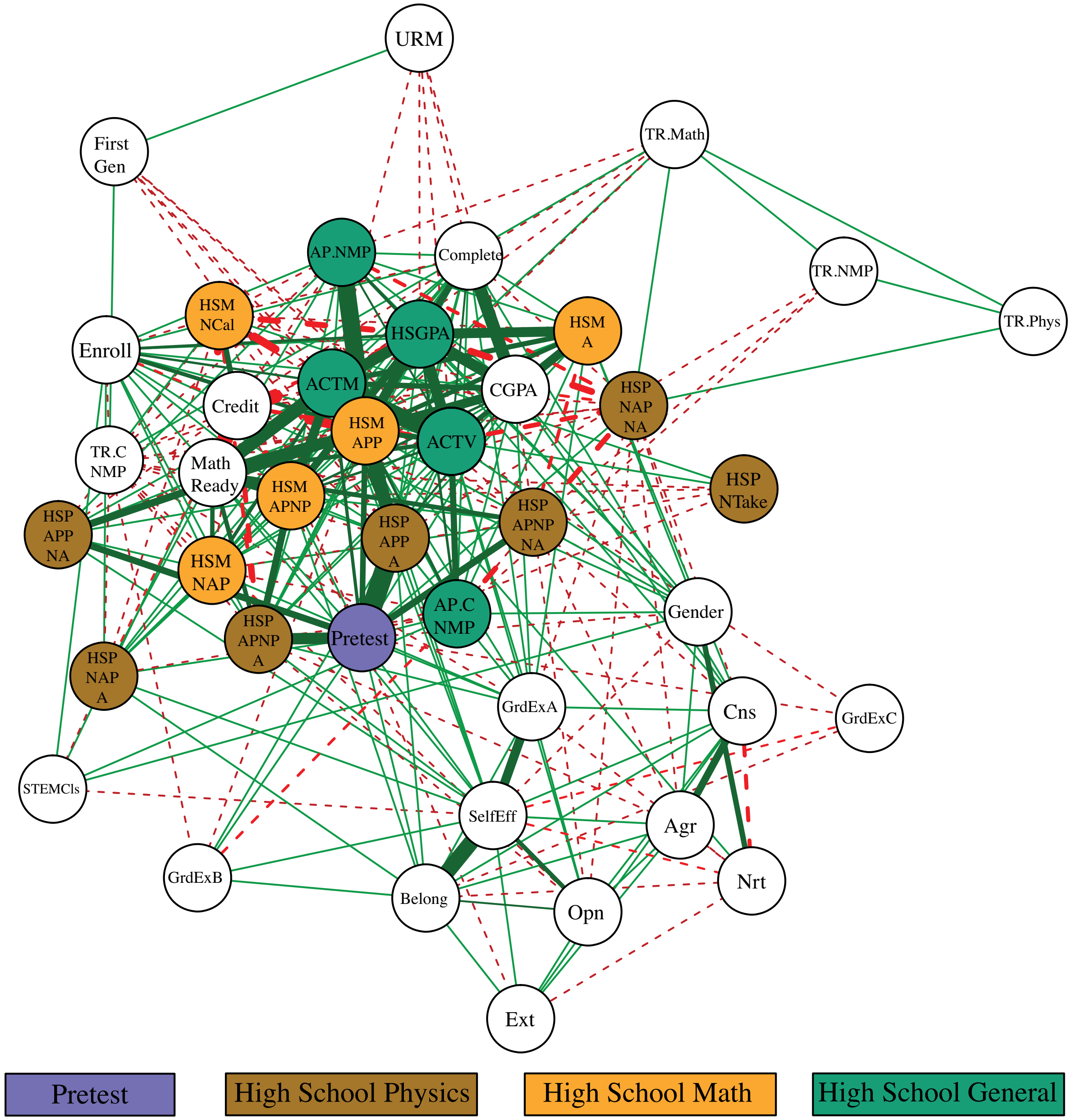

FIG. 1. Correlation matrix. Green (solid) lines represent positive correlations; red (dashed) lines negative correlations. Thicker lines represented stronger correlations.

nodes with spring constant proportional to the correlation coefficient. Repulsive Coulomb's lawlike forces are introduced between all nodes with the same effective positive charge given to all nodes. The energy of the system is them minimized pushing weakly correlated nodes away from the system and drawing strongly correlated nodes together. An alternate visualization using the "CORRPLOT" package [90] is shown in the Supplemental Material [91].

The correlation matrix helps highlight some general patterns in Tables II and III. Pretest scores are most strongly correlated with taking AP high school physics for any grade as well as taking and passing AP calculus. In general, demographic characteristics, transfer credit, and noncognitive variables are weakly related to pretest score. The noncognitive variables do share some strong relations among themselves.

C. Variable importance

In the next section, linear regression is used to build an optimal model combining all variables. The interrelations of the variables evident in the previous section raise concerns of the effect of multicolinearity on these models. There are strong theoretical reasons to believe one variable could mask the effect of another variable in a combined model, indicating it was less important than it actually was. Measures of general high school academic success such as ACT and SAT scores and high school GPA are related to general measures of college success such as college GPA. Academic success should improve self-efficacy and lead to higher grade expectations in physics classes. Specific academic success measured by course grades in high school physics and mathematics classes should be related to general academic success. Furthermore, taking an AP physics class requires the school to offer AP physics which may imply a generally more enriched academic curriculum. Students who take and pass AP calculus may be more likely to have access to AP physics.

To understand these relations, four measures of variable importance were calculated. The first uses the variable as the only independent variable in a linear regression predicting pretest score: R_f^2 measures the variance explained by this model and p_f the significance of the model (f indicates first). This model estimates the importance of the variable in exclusion of other variables. The second builds a linear model using all other variables and reports the difference in R^2 of this model and a model including the variable of interest: ΔR_l^2 measures the change in variance explained by the two models and p_l the significance of the difference (measured by an ANOVA analysis; the subscript l indicates last). This measures the additional variance explained by the variable in the presence of all other variables. This may understate the importance because of covariance with other variables. The third measure borrows a method from machine learning and measures the difference in variance explained by a model containing a randomly sampled subset of variables and a model which adds the variable of interest to the subset: the difference is captured by ΔR_s^2 and p_s (s for sampled). This is a bootstrapped method using 500 replications sampling the data with replacement which allows a standard deviation to be estimated. This measures the average importance of the variable in the presence of other variables. These three measures are calculated using the groups of variables (panels) defined in Table I; the results are presented in Table IV. The final measure attempts to predict the variable of interest using the other independent variables and reports the R^2 of this model, R_v^2. This measures how much of the information provided by the variable is available in combinations of other independent variables; this can only be performed on individual variables and not panels and is only reported in the Supplemental Material [91]. The Supplemental Material reports all four measures for each variable individually.

TABLE IV. Paneled variable importance. R_f^2 represents the variance explained when all variables in the panel are the only variables in the model. ΔR_l^2 represents the additional variance explained when all variables in the panel are added as the last variables in the model. ΔR_s^2 is the average additional variance explained when adding the variables in the panel to a model subsampling the variable list to 5 variables. p_f is the p value for the panel only model. p_l is the p value for the ANOVA test comparing the two models. The p_s value is the probability the difference ΔR_s^2 happened by chance found using a t test.

Panel	R_f^2	p_f	ΔR_l^2	p_l	ΔR_s^2	p_s
Repeat	0.000	0.664	0.011	0.000	0.007	0.000
College	0.055	0.000	0.017	0.000	0.031	0.000
Math ready	0.046	0.000	0.003	0.020	0.017	0.000
HS general	0.093	0.000	0.020	0.000	0.048	0.000
AP general	0.013	0.001	0.001	0.436	0.005	0.000
Transfer	0.004	0.342	0.001	0.831	0.005	0.000
HS physics	0.175	0.000	0.104	0.000	0.134	0.000
HS math	0.049	0.000	0.004	0.152	0.021	0.000
Belonging	0.015	0.000	0.002	0.122	0.007	0.000
Self-efficacy	0.042	0.000	0.001	0.236	0.017	0.000
Grade expectation	0.044	0.000	0.010	0.000	0.025	0.000
Personality	0.031	0.000	0.014	0.000	0.029	0.000
Demographics	0.019	0.000	0.009	0.003	0.016	0.000

Each measure of variable importance provides different information about the relation of the variable to other independent variables and to the dependent variable. The change in R^2 if the variable is the last variable added to the model, ΔR_l, is probably the most interesting because if captures how much unique variance the variable explains above all other variables. If this measure is high, the variable is capturing unique information important to the prediction of the dependent variable not captured by the other variables. This measure, however, may indicate a variable has low importance because the variable is colinear with other variables, understating its importance. The sampled variable importance, ΔR_s, partially addresses this underestimate by calculating the average additional variable explained in the presence of a randomly selected set of variables. This is probably the best estimate of the actual explanatory power of the variable. The first-in variable of importance, R_f^2, should generally overestimate the importance because some of the predictive power of the variable used as the only variable in the model comes from colinearity with other variables.

Table IV shows that high school physics taking patterns explain the greatest amount of variance in the models when used as the only variable in the model (R_f^2) or the last variable added to the model (ΔR_l^2). This panel of variables is not, however, independent of the other variables as shown by the difference in R_f^2 and ΔR_l^2; taking and doing well in high school physics is related to other more general features of academic success and access to enriched high school classes. HS general explains the second most variance when added first to the model, but little variance when added last. Differences in general high school preparation influence other variables in the model; these influences reduce the additional predictive power of this group greatly. This is also true of a number of variables explaining about 5% of the variance on their own (College, Math Ready, HS Math, Self-Efficacy, and Grade Expectation), but little variance when added to the model with the all other variables present. High school physics stands out explaining 10% additional variance when added last to the model. Beyond HS physics, only college, HS general, and personality explain at least 1% additional variance when added last to the models.

In summary, high school physics stands out as the variable that explains by far the most unique variance (10%), variance not explained by other variables, in pretest scores. Beyond high school physics, general high school preparation (HSGPA, ACT, and SAT scores) and college performance (CGPA) also explain about 2% additional variance when added as the last variable in the models.

D. Optimal pretest model

The full set of variables in Table I was used to predict the pretest score using multiple linear regression. The base level variables were removed because they are colinear with other variables in the variable's panel; the regression coefficients of the other variables in the panel measure changes with respect to the level of the base variable. This model is presented as the full pretest model in the Supplemental Material [91]. An optimal model, also shown in the Supplemental Material, was constructed by removing dependent variables that were not statistically significant at the $p < 0.05$ level. All variables representing the coding of a multilevel categorical variable such as HS physics were retained if one of the variables was significant. The optimal model was statistically equivalent to the full model [$F(15, 1079) = 0.97$, $p = 0.48$] using ANOVA and explained $R^2 = 0.31$ of the variance in pretest score.

The original model contained 36 independent variables, therefore, construction of the optimal model involved performing 36 statistical tests. If a Bonferroni correction is applied to the $p < 0.05$ significance level to correct for the number of statistical tests, the new significant threshold is $p < 0.05/36 = 0.0014$. A Bonferroni corrected optimal model removing variables that did not meet the corrected significance level was constructed and is presented Table V. The corrected model was statistically inferior to the optimal model, [$F(7, 1094) = 6.83$, $p = 0.00000$] and explained $R^2 = 0.28$ of the variance. A Bonferroni correction removes significant regressors and therefore lowers R^2. The corrected model is used in future discussion; interested readers can consult the Supplemental Material [91] for the full model.

TABLE V. Optimal pretest model with Bonferroni correction. B is the regression coefficient, SE the standard error, β the standardized regression coefficient, t the t statistic, and p the probability a value as large or larger than t occurred by chance. The overall model explains $R^2 = 0.28$ [$F(14, 1101) = 30.4$, $p = 0.00000$] of the variance in pretest score.

	B	SE	β	t	p
(Intercept)	6.34	5.70	−0.53	1.11	0.26621
HSP.APP.A	33.22	2.97	1.82	11.17	0.00000
HSP.APNP.A	16.80	1.64	0.92	10.22	0.00000
ACTV	0.21	0.03	0.20	6.82	0.00000
HSP.APNP.NA	11.46	1.73	0.63	6.62	0.00000
Gender	−5.64	1.07	−0.31	−5.29	0.00000
GrdExA	6.25	1.38	0.34	4.55	0.00001
HSP.NAP.A	5.59	1.31	0.31	4.27	0.00002
HSP.APP.NA	21.52	5.06	1.18	4.25	0.00002
CGPA	4.66	1.17	0.12	3.99	0.00007
Repeat	8.15	2.25	0.45	3.62	0.00031
Ext	−2.28	0.64	−0.09	−3.58	0.00036
HSGPA	−4.28	1.33	−0.10	−3.21	0.00136
HSP.NAP.NA	4.97	1.57	0.27	3.17	0.00156
GrdExB	2.38	1.34	0.13	1.77	0.07721

Table V presents both the unstandardized regression coefficient B, its standard error SE, and the standardized regression coefficient β. The standardized coefficient is calculated by repeating the regression with all continuous variables normalized by subtracting the mean and dividing by the standard deviation. For dichotomous independent variables, B measures the difference in the percentage pretest score between the two levels of the dichotomous variable and β measures the difference in the normalized pretest scores between the two levels of the variable in standard deviation units; as such, it can be interpreted as an effect size using Cohen's criteria for the d statistic. For continuous independent variables, B measures the change in the pretest percentage score when the independent variable is increased by one unit; β represents the change in pretest scores in standard deviation units for a 1 standard deviation change in the independent variable. β also represents the correlation between the independent and dependent variables (correcting for other variables) and may be interpreted using Cohen's effect size criteria for r. For the dichotomous variables, repeating the class was near a medium effect as was taking AP physics, receiving a grade less than A, and not passing the AP test. Passing the AP physics test, as well as not passing but earning an A in the AP class, were all large effects. For the continuous variables, no variable produced more than a small effect.

In summary, the optimal pretest regression model explained 28% of the variance in pretest score. As might be expected from the variable importance analysis, high school physics variables had by far the largest regression coefficients. Interestingly, repeating the class produced a fairly large positive effect when other factors were controlled for. A student's belief that they would do well and earn an A in the class was also important, showing some noncognitive factors remain important even after controlling for academic factors.

E. Optimal post-test model

The same set of variables now including pretest score was used to predict post-test score. The full regression model is presented in the Supplemental Material [91]. An optimal model was constructed by removing independent variables which were not significant at the $p < 0.05$ level. This model is also presented in the Supplemental Material. This model explained $R^2 = 0.56$ of the variance in post-test score. This model was not statistically different from the full model using an ANOVA test [$\mathrm{F}(21, 1078) = 1.39$, $p = 0.1103$]. As before, a Bonferroni corrected model was constructed, which was statistically inferior to the optimal model, [$\mathrm{F}(9, 1099) = 5.42$, $p = 0.00000$], but explained 54% of the variance. We will focus on this model presented in Table VI as it explains the majority of the variance and contains only the variables most important to predicting the post-test score.

The optimal model for the post-test was dramatically different from the model for the pretest; the post-test model was missing the variables related to high school physics. This seems to indicate that the pretest score fully captures the effects of high school physics and that there are not additional effects of high school physics on the post-test score. To further test this conclusion, the post-test models were fit without using pretest score as an independent variable. The models are shown in the Supplemental Material [91]. Without pretest score in the model, high school physics is a strong predictor of post-test score. As such, there seems little advantage in conceptual learning conferred by taking high school physics that is not measured by the pretest. The advantages conferred by relearning the material are not important in the overall conceptual understanding developed in the class.

The Bonferroni corrected optimal model (Table VI) contains 6 variables (grade expectation is a single categorical variable) and explains 54% of the variance in post-test score; some variables are generally unavailable to physics faculty such as self-efficacy or grade expectation. A simplified model containing only pretest score, gender, and ACT and SAT scores explained 53% of the variance and is shown in Table VII. One can then progressively remove variables to determine how much variance each explains. Removing gender produced a model which explained 50% of the variance (Table VIII); gender explained 3% additional variance controlling for pretest score and ACT and SAT scores. Removing ACT and SAT mathematics and verbal percentile scores produced a model which explained 44% of the variance using only the pretest score. Pretest alone explained 44% of the variance in post-test score; ACT and SAT scores explained an additional 6% of the variance controlling for pretest score. This should not imply these variables are independent; a model with ACT and SAT scores but not pretest scores explains 18% of the variance in post-test scores.

In summary, the optimal model explained 54% of the variance in post-test scores. By far the most important variable in the model was the pretest score explaining 44% of the variance on its own. ACT and SAT scores (6%) and gender (3%) were also important in the model. The high school physics variables were no longer significant predictors when pretest was controlled for.

TABLE VI. Optimal post-test model with Bonferroni correction. B is the regression coefficient, SE the standard error, β the standardized regression coefficient, t the t statistic, and p the probability a value as large or larger than t occurred by chance. The overall model explains $R^2 = 0.54$ [$\mathrm{F}(7, 1108) = 184.25$, $p = 0.00000$] of the variance in post-test score.

	B	SE	β	t	p
(Intercept)	−22.45	4.67	0.27	−4.81	0.00000
Pretest	0.85	0.03	0.56	25.34	0.00000
Gender	−10.13	1.30	−0.37	−7.79	0.00000
ACTM	0.31	0.05	0.16	6.19	0.00000
ACTV	0.23	0.04	0.15	5.71	0.00000
SelfEff	3.47	0.87	0.09	3.97	0.00008
GrdExB	−5.78	1.64	−0.21	−3.53	0.00044
GrdExA	−5.32	1.73	−0.19	−3.07	0.00217

TABLE VII. Post-test model with pretest, gender, and ACT and SAT scores. B is the regression coefficient, SE the standard error, β the standardized regression coefficient, t the t statistic, and p the probability a value as large or larger than t occurred by chance. The overall model explains $R^2 = 0.53$ [$\mathrm{F}(4, 1111) = 310.03$, $p = 0.00000$] of the variance in the post-test score.

	B	SE	β	t	p
(Intercept)	−13.45	3.46	0.11	−3.89	0.00011
Pretest	0.86	0.03	0.56	25.89	0.00000
ACTM	0.33	0.05	0.17	6.48	0.00000
ACTV	0.22	0.04	0.14	5.38	0.00000
Gender	−10.69	1.30	−0.39	−8.24	0.00000

TABLE VIII. Post-test model with pretest and ACT and SAT scores. B is the regression coefficient, SE the standard error, β the standardized regression coefficient, t the t statistic, and p the probability a value as large or larger than t occurred by chance. The overall model explains $R^2 = 0.50$ [$\mathrm{F}(3, 1112) = 368.56$, $p = 0.00000$] of the variance in the post-test score.

	B	SE	β	t	p
(Intercept)	−15.33	3.55	−0.00	−4.32	0.00002
Pretest	0.90	0.03	0.59	26.69	0.00000
ACTM	0.37	0.05	0.18	6.99	0.00000
ACTV	0.15	0.04	0.10	3.67	0.00025

IV. DISCUSSION

This study investigated three research questions which will be discussed in the order proposed. Many results were discussed as they were presented, and therefore, the following summarizes the findings.

RQ1: What academic and noncognitive factors are most important in predicting FMCE pretest scores? This work found that high school physics preparation was a very important feature in explaining pretest score. High school physics has been investigated in other works with varying outcomes [30–33]. This work illustrates the importance of the details of the student's high school experience captured by whether the course was AP and their performance in the high school class captured by their grade. The variable HSP.NTake captures whether the student took any high school physics; this variable explains only 4.6% of the variance in pretest score. The difference in pretest score between students who took high school physics and those who did not was 9%, a medium effect ($d = 0.52$). This was a larger effect size than the difference observed in normalized gain scores by Hake [30], but commensurate to the difference in pretest scores reported by Kost *et al.* [33]. The difference between students who took no high school physics and those who had high school physics changed dramatically with the type of high school physics and whether the student earned an A in the high school physics class. A student who took AP physics, passed the AP test, and earned an A had a pretest score 37% percentage points higher than a student with no high school physics, an extremely large effect ($d = 2.7$). The full set of variables in the HS physics panel which capture both the kind of high school physics taken and the student's grade explained 17.5% of the variance in pretest score, the majority of the variance explained by the full set of variables (31%, 28% if Bonferroni corrected).

Many other variables had correlations with pretest score above the threshold of a small effect as shown in Table II. Beyond high school physics, ACT and SAT scores, being math ready, and the student's expected grade in the physics class had substantial correlations. Many dichotomous variables showed substantial differences in pretest score between the two levels of the variable (Table III). Again beyond high school physics, being math ready, expecting an A in the physics class, and passing the AP calculus test were at or near medium effects. Correlation analysis supported the central importance of taking high school physics and the type of high school physics taken in predicting pretest scores, Fig. 1.

Measures of variable importance, Table IV, continued to support the central role high school physics preparation plays in pretest scores with the HS Physics group of variables explaining 10% additional variance when added to a model containing all other variables, five times as much additional variance as any other group of variables. High school preparation is not the only factor affecting pretest scores controlling for other variables; college and general high school academic achievement explained 2% additional variance and personality 1% additional variance.

Bonferroni corrected linear regression analysis, Table V, also supported the centrality of high school physics in predicting pretest score, but also the role of some additional variables. Variables in the HS Physics group had β coefficients with the largest effect sizes with many above the threshold of 0.5 for a medium effect. Repeating the class was near a medium effect. Believing one would receive an A in the class and gender were small effects as was college CGPA and ACT and SAT verbal score.

Overall, only 28% of the variance in pretest scores was explained by the extensive set of variables used in this study. There are many potential reasons for the amount of variance unexplained. First, pretest scores were on average fairly low (the pretest average was 23%). For this student population, the FMCE is not well calibrated to discriminate different incoming levels of conceptual understanding (the pretest is too difficult for these students). With Thornton scoring [92], the total FMCE score is 33 and each item contributes approximately 3.3% to the total score. The 10% difference in pretest between taking and not taking high school physics is then only 3 items. A single misconception addressed in one high school classroom, but not addressed in another, could account for this difference (the FMCE presents many items testing the same misconception). There are potentially broad variations of the high school experience of students in each level of the high school physics variable which could affect pretest scores.

RQ2: What academic and noncognitive factors are most important in predicting FMCE post-test scores correcting for FMCE pretest scores? The Bonferroni corrected optimal post-test linear regression model, Table VI, which contained the pretest score as a variable did not contain any high school physics variables; pretest score fully controlled for the effect of high school physics preparation. Therefore, there is not an additional effect of relearning material that is evident on the post-test. If the pretest was inaccurately measuring the student's prior preparation because they had physics knowledge they had forgotten, we would expect prior high school preparation variables to affect the post-test in ways not controlled for by the pretest. This suggests the student does not have hidden conceptual knowledge at the time of the pretest that he or she knew at a previous time but has forgotten. We note this result does not imply there are not benefits of relearning forgotten physics knowledge that impact other factors in a physics classes such as quantitative or procedural knowledge.

Post-test scores also depended on general high school academic preparation measured by ACT and SAT mathematics and verbal scores and a student's belief in their ability to succeed in the class measured both the self-efficacy and by their grade expectation. The optimal model explained 54% of the variance in post-test score, which is substantial but far from perfect.

The variables grade expectation and self-efficacy were collected through a survey instrument and would not be available to most instructors. Removing these variables reduced the variance explained by the model by only 1%. The primary variables predicting post-test score were pretest score (44% of the variance alone), ACT and SAT scores, an additional 6% of the variance when added to a model containing pretest score (18% of the variance on its own), and gender explaining 3% of the variance when added to a model containing pretest score and ACT and SAT scores. As such, the majority of the variance in post-test score is explained by pretest scores, but some prior academic achievement variables and demographic variables are also important.

RQ3: Do academic and noncognitive factors explain gender differences in FMCE pretest and post-test scores? Part of the motivation for this work was to determine if gender differences in pretest and post-test scores could be explained by differences in high school preparation or differences in noncognitive factors. The overall gender difference in pretest score was 5.1%; the difference grew to 13.8% on the post-test. The models controlling for both high school preparation and noncognitive factors (Tables V and VI) failed to account for the gender difference. If noncognitive or high school preparation were the source of the gender difference, the gender regression coefficient would have been reduced in these models (these factors would have mediated the gender difference). This was not the case for the pretest where the gender regression coefficient in the model presented in Table V showed a gender difference of 5.6% correcting for all these factors. The gender coefficient in the post-test model was reduced slightly to 10.1%, but most of the original gender difference remained unexplained. As such, none of the gender difference in pretest scores was explained by noncognitive factors or high school preparation while $(13.8\% - 10.13\%)/13.8\% = 27\%$ of the difference in post-test scores was explained by these factors and pretest scores. Thus, neither noncognitive nor high school preparation differences account for the majority of the gender difference in either pretest or post-test scores at the institution studied. This observation does not support Salehi's *et al.* [1] finding that prior preparation variables fully mediated gender differences in final examination scores. It is consistent with Stewart's *et al.* observation the much of the gender difference in post-test scores are unexplained by the same factors [2].

V. IMPLICATIONS

This study examined the features predicting pretest scores with a large sample and an extensive set of high school level, college level, and noncognitive variables. The total variance explained with all these measures was only 28%. As such, an instructor using pretest scores should anticipate that there is some, possibly substantial, uncertainty in the pretest scores of individual students. If pretest scores are used for instructional decisions, they should be used cautiously. Likewise, uncertainly in pretest scores will generate uncertainty in the absolute gain and the normalized gain.

Because of this uncertainly, small differences in pretest scores are not particularly meaningful and only large differences in scores are potentially useful for instructional decisions. The high school physics panel in Table III can provide some guidance in these decisions. Students who have not had high school physics may need additional support; the average pretest score of these students is 16%. Students who had high school physics, but not AP physics, and were successful earning an A scored on average 22% on the pretest. A student who had the enriched AP curriculum, earned an A but did not pass the AP test, scored on average 34%; those who pass the AP test 53%. As such, 10% differences in pretest score imply a substantially different high school physics experience.

This work showed that by far the most important variable in predicting pretest scores was the type of high school physics and the student's grade in high school physics which predicted 17.5% of the variance in pretest score alone. The kind of high school physics (whether or not it was AP physics) and how the student did in the physics class and on the AP test was crucial to the predictive power of high school physics. Whether or not the student took any kind of high school physics explained only 4.6% of the variance. As such, researchers seeking to explore the role of high school physics preparation should gather detailed information about the high school experience.

For the optimal pretest regression model, Table V, measures of general high school academic and college achievement were important, but not as important as high school preparation. As such, a pretest score measures primarily prior preparation in physics but is also influenced by the general high school academic preparation and college academic achievement of the student. This means that pretest scores may change with factors not related to specific preparation in physics which confounds their use as a control for prior knowledge of physics.

For the optimal post-test regression model, Table VI, the majority of the variance was explained by pretest scores (44%), ACT and SAT scores explained an additional 6% of the variance, and gender 3%. All other variables only explained 1% together. As such, pretest score captures most of the effect of prior preparation and noncognitive effects on post-test score and should act as a good, but not perfect, control for these effects.

The relation of post-test scores to pretest scores and other variables have important implications for an ongoing debate in PER about how to measure conceptual learning gains in a physics class [22]; specifically how can and should the normalized gain popularized by Hake [10] be updated? Nissen *et al.* showed the normalized gain was biased toward students with higher pretest scores [22]. This work showed that both pretest and post-test scores were related to general high school level achievement measured by ACT and SAT scores and that pretest score was also related to general college level achievement measured by college GPA; these relations should also bias normalized gain toward populations with higher scores on these measures. A substantial number of studies suggest some groups underrepresented in physics have lower general high school achievement scores than their majority peers [1,2,93]. These differences are partially the result of limited access to advanced coursework in schools serving underrepresented students [94].

Examining the differences of pretest scores by level of high school preparation in Table III shows that there is a broad spectrum of prior preparation in physics in the class studied. It is important to take this into consideration when designing activities in the class and interpreting the results of assessment so that all students in the class can have the chance to succeed. If only a small subset of students seem to grasp some part of the material, it may be because they understood it before starting the class.

This work identified taking, but possibly not passing, an AP physics course as an important factor in predicting FMCE pretest scores in college physics classes. The focus on AP was not meant to suggest that other enriched curriculum such as the International Baccalaureate (IB) program could not produce similar results. There were insufficient students in these programs in the sample to draw statistical conclusions. It was, however, clear that classes taken in college during high school, transfer classes, were of little benefit in producing conceptual understanding in college. It is unclear if this is because of the variable quality and content of these courses, or because they are often offered by school districts unable to make the investment required to offer AP physics classes. This lack of resources could have general negative effects on the academic program.

VI. LIMITATIONS

This work was performed at a single institution. The work should be replicated at institutions with a range of incoming students levels of preparation and demographic composition so as to determine if the results are general.

VII. CONCLUSION

This study applied correlation analysis and linear regression analysis to understand the relation of high school preparation, college achievement, and noncognitive factors to students' physics conceptual understanding measured by the FMCE and whether any of these factors explained gender differences in FMCE pretest and post-test scores.

Several academic and noncognitive factors were significant in predicting FMCE pretest scores including high school physics preparation, high school math preparation, ACT and SAT verbal scores, college GPA, and the student's expected grade. Whether the student had taken high school physics explained 4.6% percent of the variance in pretest score. The kind of high school physics (normal or AP) and the student's grade in high school physics explained substantially more variance, 17.5%. ACT or SAT verbal and mathematics scores, students' grade expectation and self-efficacy were significant in predicting post-test score while controlling for pretest scores. High school physics taking patterns were not important in predicting post-test scores if pretest scores were controlled for. As such, pretest scores completely captured the effects of high school preparation on post-test scores in the regression analyses. Gender differences observed in FMCE pretest and post-test scores were changed little by controlling for either high school preparation or noncognitive factors.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation under Grants No. ECR-1561517 and No. HRD-1834569.

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

[2] J. Stewart, G. L. Cochran, R. Henderson, C. Zabriskie, S. DeVore, P. Miller, G. Stewart, and L. Michaluk, Mediational effect of prior preparation on performance differences of students underrepresented in physics, Phys. Rev. Phys. Educ. Res. **17**, 010107 (2021).

[3] A. Madsen, S. B. McKagan, and E. 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).

[4] 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).

[5] R. Henderson, C. Zabriskie, and J. Stewart, Rural and first generation performance differences on the Force and Motion Conceptual Evaluation, presented at PER Conf. 2018, College Park, MD, 10.1119/perc.2018.pr.Henderson.

[6] J. M. Nissen and J. T. Shemwell, Gender, experience, and self-efficacy in introductory physics, Phys. Rev. Phys. Educ. Res. **12**, 020105 (2016).

[7] I. A. Halloun and D. Hestenes, The initial knowledge state of college physics students, Am. J. Phys. **53**, 1043 (1985).

[8] I. A. Halloun and D. Hestenes, Common sense concepts about motion, Am. J. Phys. **53**, 1056 (1985).

[9] D. Hestenes, M. Wells, and G. Swackhamer, Force concept inventory, Phys. Teach. **30**, 141 (1992).

[10] 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).

[11] S. Freeman, S. L. Eddy, M. McDonough, M. K. Smith, N. Okoroafor, H. Jordt, and M. P. Wenderoth, Active learning increases student performance in science, engineering, and mathematics, Proc. Natl. Acad. Sci. U.S.A. **111**, 8410 (2014).

[12] R. K. Thornton and D. R. Sokoloff, Assessing student learning of Newton's laws: The Force and Motion Conceptual Evaluation and the evaluation of active learning laboratory and lecture curricula, Am. J. Phys. **66**, 338 (1998).

[13] D. P. Maloney, T. L. O'Kuma, C. Hieggelke, and A. Van Huevelen, Surveying students' conceptual knowledge of electricity and magnetism, Phys. Educ. Res., Am. J. Phys. Suppl. **69**, S12 (2001).

[14] R. Chabay and B. Sherwood, Qualitative understanding and retention, AAPT Announcer **27**, 96 (1997).

[15] Physport, https://www.physport.org.

[16] L. C. McDermott and E. F. Redish, Resource letter: PER-1: Physics education research, Am. J. Phys. **67**, 755 (1999).

[17] J. L. Docktor and J. P. Mestre, Synthesis of discipline-based education research in physics, Phys. Rev. Phys. Educ. Res. **10**, 020119 (2014).

[18] A. Madsen, S. B. McKagan, and E. C. Sayre, Resource letter RBAI-1: Research-based assessment instruments in physics and astronomy, Am. J. Phys. **85**, 245 (2017).

[19] D. E. Meltzer and R. K. Thornton, Resource letter ALIP-1: Active-learning instruction in physics, Am. J. Phys. **80**, 478 (2012).

[20] C. M. Schroeder, T. P. Scott, H. Tolson, T. Y. Huang, and Y. H. Lee, A meta-analysis of national research: Effects of teaching strategies on student achievement in science in the United States, J. Res. Sci. Teach. **44**, 1436 (2007).

[21] P. L. Bonate, *Analysis of Pretest-Posttest Designs* (CRC Press, Monterey, CA, 2000).

[22] J. M. Nissen, R. M. Talbot, A. N. Thompson, and B. Van Dusen, Comparison of normalized gain and Cohen's *d* for analyzing gains on concept inventories, Phys. Rev. Phys. Educ. Res. **14**, 010115 (2018).

[23] D. S. Hewagallage, J. Stewart, and R. Henderson, Differences in the predictive power of pretest scores of students underrepresented in physics, presented at PER Conf. 2019, Provo, UT, 10.1119/perc.2019.pr.Hewagallage.

[24] D. Liberman and H. T. Hudson, Correlation between logical abilities and success in physics, Am. J. Phys. **47**, 784 (1979).

[25] H. T. Hudson and D. Liberman, The combined effect of mathematics skills and formal operational reasoning on student performance in the general physics course, Am. J. Phys. **50**, 1117 (1982).

[26] A. B. Champagne, L. E. Klopfer, and J. H. Anderson, Factors influencing the learning of classical mechanics, Am. J. Phys. **48**, 1074 (1980).

[27] D. E. Meltzer, The relationship between mathematics preparation and conceptual learning gains in physics: A possible "hidden variable" in diagnostic pretest scores, Am. J. Phys **70**, 1259 (2002).

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

[29] V. P. Coletta, J. A. Phillips, and J. J. Steinert, Interpreting Force Concept Inventory scores: Normalized gain and SAT scores, Phys. Rev. Phys. Educ. Res. **3**, 010106 (2007).

[30] 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 (unpublished).

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

[32] 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).

[33] L. E. Kost, S. J. Pollock, and N. D. Finkelstein, Characterizing the gender gap in introductory physics, Phys. Rev. Phys. Educ. Res. **5**, 010101 (2009).

[34] S. Ramlo, Validity and reliability of the Force and Motion Conceptual Evaluation, Am. J. Phys. **76**, 882 (2008).

[35] J. Yang, C. Zabriskie, and J. Stewart, Multidimensional item response theory and the Force and Motion Conceptual Evaluation, Phys. Rev. Phys. Educ. Res. **15**, 020141 (2019).

[36] J. Wells, R. Henderson, A. Traxler, P. Miller, and J. Stewart, Exploring the structure of misconceptions in the Force and Motion Conceptual Evaluation with modified module analysis, Phys. Rev. Phys. Educ. Res. **16**, 010121 (2020).

[37] 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).

[38] T. I. Smith, M. C. Wittmann, and T. Carter, Applying model analysis to a resource-based analysis of the Force and Motion Conceptual Evaluation, Phys. Rev. Phys. Educ. Res. **10**, 020102 (2014).

[39] T. I. Smith and M. C. Wittmann, Applying a resources framework to analysis of the Force and Motion Conceptual Evaluation, Phys. Rev. Phys. Educ. Res. **4**, 020101 (2008).

[40] L. Crocker and J. Algina, *Introduction to Classical and Modern Test Theory* (Holt, Rinehart, and Winston, Mason, OH, 1986).

[41] 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).

[42] P. A. Westrick, H. Le, S. B. Robbins, J. M. R. Radunzel, and F. L. Schmidt, College performance and retention: A meta-analysis of the predictive validities of ACT scores, high school grades, and SES, Educ. Assess. **20**, 23 (2015).

[43] M. C. Frey and D. K. Detterman, Scholastic assessment or g? The relationship between the scholastic assessment test and general cognitive ability, Psychol. Sci. **15**, 373 (2004).

[44] K. A. Koenig, M. C. Frey, and D. K. Detterman, ACT and general cognitive ability, Intelligence **36**, 153 (2008).

[45] P. A. Westrick, J. P. Marini, L. Young, H. Ng, D. Shmueli, and E. J. Shaw, *Validity of the SAT for predicting first-year grades and retention to the second year* (College Board, New York, NY, 2019).

[46] J. Allen, *Updating the ACT College Readiness Benchmarks. ACT Research Report Series 2013 (6).* (ACT, Inc., Iowa City, IA, 2013).

[47] W. J. Camara and G. Echternacht, The SAT I and high school grades: Utility in predicting success in college, Res. Not. **10**, 3 (2000), https://eric.ed.gov/?id=ED446592.

[48] S. Geiser and R. Studley, UC and the SAT: Predictive validity and differential impact of the SAT I and SAT II at the University of California, Educ. Assess. **8**, 1 (2002).

[49] S. Geiser and M. V. Santelices, *Validity of High-School Grades in Predicting Student Success beyond the Freshman Year: High-School Record vs. Standardized Tests as Indicators of Four-Year College Outcomes. Research & Occasional Paper Series: CSHE. 6.07.* (Center for Studies in Higher Education, University of California-Berkeley, Berkeley, CA, 2007).

[50] W. G. Bowen, W. M. Chingos, and M. S. McPherson, Test scores and high school grades as predictors, *Crossing the Finish Line* (Princeton University Press, Princeton, NJ, 2009), p. 112.

[51] B. M. Galla, E. P. Shulman, B. D. Plummer, M. Gardner, S. J. Hutt, J. P. Goyer, S. K. D'Mello, A. S. Finn, and A. L. Duckworth, Why high school grades are better predictors of on-time college graduation than are admissions test scores: The roles of self-regulation and cognitive ability, Am. Educ. Res. J. **56**, 2077 (2019).

[52] M. Richardson, C. Abraham, and R. Bond, Psychological correlates of university students' academic performance: A systematic review and meta-analysis, Psychol. Bull. **138**, 353 (2012).

[53] A. E. Poropat, A meta-analysis of the five-factor model of personality and academic performance, Psychol. Bull. **135**, 322 (2009).

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

[55] S. Andrew, Self-efficacy as a predictor of academic performance in science, J. Adv. Nurs. **27**, 596 (1998).

[56] S. Lau and R. W. Roeser, Cognitive abilities and motivational processes in high school students' situational engagement and achievement in science, Educ. Assessment **8**, 139 (2002).

[57] W. J. Hughes, Perceived gender interaction and course confidence among undergraduate science, mathematics, and technology majors, J. Women Minor. Sci. Engineer. **6**, 1 (2000).

[58] 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).

[59] S. L. Eddy and S. E. Brownell, Beneath the numbers: A review of gender disparities in undergraduate education across science, technology, engineering, and math disciplines, Phys. Rev. Phys. Educ. Res. **12**, 020106 (2016).

[60] M. E. Junge and B. J. Dretzke, Mathematical self-efficacy gender differences in gifted/talented adolescents, Gifted Child Quar. **39**, 22 (1995).

[61] M. Besterfield-Sacre, M. Moreno, L. J. Shuman, and C. J. Atman, Gender and ethnicity differences in freshmen engineering student attitudes: A cross-institutional study, J. Engin. Educ. **90**, 477 (2001).

[62] R. Dou, E. Brewe, J. P. Zwolak, G. Potvin, E. A. Williams, and L. H. Kramer, Beyond performance metrics: Examining a decrease in students' physics self-efficacy through a social networks lens, Phys. Rev. Phys. Educ. Res. **12**, 020124 (2016).

[63] V. Sawtelle, E. Brewe, and L. H. Kramer, Positive impacts of modeling instruction on self-efficacy, AIP Conf. Proc. **1289**, 289 (2010).

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

[65] B. M. K. Hagerty, J. Lynch-Sauer, K. L. Patusky, M. Bouwsema, and P. Collier, Sense of belonging: A vital mental health concept, Arch. Psychiatric Nurs. **6**, 172 (1992).

[66] 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).

[67] 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).

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

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

[70] L. R. Goldberg, The development of markers for the big-five factor structure, Psychol. Assess. **4**, 26 (1992).

[71] O. P. John, E. M. Donahue, and R. L. Kentle, *The Big Five Inventory–Versions 4a and 54* (Institute of Personality and Social Research, University of California-Berkeley,, Berkeley, CA, 1991).

[72] O. P. John, L. P. Naumann, and C. J. Soto, Paradigm shift to the integrative big five trait taxonomy: History,

measurement, and conceptual issues, in *Handbook of Personality: Theory and Research* (The Guilford Press, New York, NY, 2008), p. 114.

[73] J. Stewart and D. Hewagallage, The relation of personality, gender, and achievement in science classes, in *American Educational Research Association Conference Proceedings, Toronto, Ontario* (American Educational Research Association, New York, NY, 2019).

[74] N. M. Else-Quest, J. S. Hyde, and M. C. Linn, Cross-national patterns of gender differences in mathematics: A meta-analysis, Psychol. Bull. **136**, 103 (2010).

[75] X. Ma, A meta-analysis of the relationship between anxiety toward mathematics and achievement in mathematics, J. Res. Math. Educ. **30**, 520 (1999).

[76] J. V. Mallow and S. L. Greenburg, Science anxiety: Causes and remedies, J. Coll. Sci. Teach. **11**, 356 (1982).

[77] M. K. Udo, G. P. Ramsey, and J. V. Mallow, Science anxiety and gender in students taking general education science courses, J. Sci. Educ. Technol. **13**, 435 (2004).

[78] J. Mallow, H. Kastrup, F. B. Bryant, N. Hislop, R. Shefner, and M. Udo, Science anxiety, science attitudes, and gender: Interviews from a binational study, J. Sci. Educ. Technol. **19**, 356 (2010).

[79] J. R. Shapiro and A. M. Williams, The role of stereotype threats in undermining girls' and women's performance and interest in STEM fields, Sex Roles **66**, 175 (2012).

[80] R. K. Thornton, D. Kuhl, K. Cummings, and J. Marx, Comparing the Force and Motion Conceptual Evaluation and the force concept iinventory, Phys. Rev. Phys. Educ. Res. **5**, 010105 (2009).

[81] US News and World Report: Education, US News and World Report, Washington, DC, https://premium.usnews.com/best-colleges. Accessed 2/23/2019.

[82] E. Mazur, *Peer Instruction: A User's Manual* (Prentice Hall, Upper Saddle River, NJ, 1997).

[83] D. Van der Linden, J. te Nijenhuis, and A. B. Bakker, The general factor of personality: A meta-analysis of Big Five intercorrelations and a criterion-related validity study, Journal of research in personality **44**, 315 (2010).

[84] P. R. Pintrich, D. A. F. Smith, T. Garcia, and W. J. Mckeachie, Reliability and predictive validity of the motivated strategies for learning questionnaire (MSLQ), Educ. Psychol. Meas. **53**, 801 (1993).

[85] T. G. Duncan and W. J. McKeachie, The making of the motivated strategies for learning questionnaire, Educ. Psych. **40**, 117 (2005).

[86] C. M. Vogt, D. Hocevar, and L. S. Hagedorn, A social cognitive construct validation: Determining women's and men's success in engineering programs, J. High. Educ. **78**, 337 (2007).

[87] R Core Team, R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing, Vienna, Austria, 2017).

[88] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences* (Academic Press, New York, NY, 1977).

[89] S. Epskamp, A. O. J. Cramer, L. J. Waldorp, V. D. Schmittmann, and D. Borsboom, qgraph: Network visualizations of relationships in psychometric data, J. Stat. Softw. **48**, 1 (2012).

[90] T. Wei and V. Simko, R package corrplot: Visualization of a correlation matrix (2021) (Version 0.90).

[91] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevPhysEducRes.18.010149 for an alternate visualization of the correlation matrix and additional regression models.

[92] R. K. Thornton and D. R. Sokoloff, Assessing student learning of Newton's laws: The Force and Motion Conceptual Evaluation and the evaluation of active learning laboratory and lecture curricula, Am. J. Phys. **66**, 338 (1998).

[93] The ACT Profile Report—National Graduating Class 2016, ACT Inc., Iowa City, IA (2016).

[94] P. R. Aschbacher, E. Li, and E. J. Roth, Is science me? High school students' identities, participation and aspirations in science, engineering, and medicine, J. Res. Sci. Teach. **47**, 564 (2010).