2023

Examining the Relations among Academic and Non-Cognitive Factors and Student Achievement

Dona Sachini Hasanjalie Hewagallage
West Virginia University, dhh0001@mix.wvu.edu

Examining the Relation among Academic and Non-Cognitive Factors and Student Achievement

Dona S. H. Hewagallage

Dissertation submitted
to the Eberly College of Arts and Sciences
at West Virginia University

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in
Physics

John Stewart, Ph.D., Chair
Paul Miller, Ph.D.
Weichao Tu, Ph.D.
Jake Follmer, Ph.D.

Department of Physics and Astronomy

Morgantown, West Virginia
2023

Keywords: physics, education, conceptual inventories, achievement, regression, correlation

ABSTRACT

Examining the Relations among Academic and Non-Cognitive Factors and Student Achievement

Dona S. H. Hewagallage

Abstract

Since the 1980s, Physics Education Research (PER) has explored the factors influencing students' success in college. This manuscript reports three different studies to understand the impact of different factors on students' college physics achievement.

The first study explored several academic (high school physics preparation, high school preparation, math readiness, and ACT or SAT verbal and mathematics scores) and non-cognitive (self-efficacy, personality, belonging, grade expectation, and demographic) factors using correlation and linear regression analysis to understand their relation to students' physics conceptual understanding measured by the Force and Motion Conceptual Evaluation (FMCE). High school preparation was found to be the most important factor in predicting conceptual understanding; the type of physics classes taken in high school, the performance in those classes, students' self-efficacy, and their grade expectations had a substantial relation to conceptual understanding.

The second study investigated two factor structures suggested by the Adams *et al.* [1] and Douglas *et al.* [2] for the Colorado Learning Attitudes about Science Survey (CLASS) instrument using correlation analysis, Exploratory Factor Analysis (EFA), and Confirmatory Factor Analysis (CFA). A new subscale model for the instrument based was proposed. The original eight-factor model of Adams *et al.* [1] could not be supported by factor analysis. The factor structure suggested by Douglas *et al.* [2] did not have good model fit parameters. A four-subscale model was developed to provide the good model fit.

The third study investigated the relations between the five-factor model of personality (agreeableness, conscientiousness, extraversion, neuroticism, and openness), self-efficacy toward physics and mathematics, and course outcomes in university physics and mathematics classes. Women reported significantly higher neuroticism in all classes and significantly higher conscientiousness in some classes while men reported higher self-efficacy. Conscientiousness and neuroticism mediated the relation of gender to self-efficacy. Self-efficacy mediated the relation of conscientiousness to course grades in all classes.

DEDICATION

To my father,
who shared this dream with me.

ACKNOWLEDGMENTS

First and foremost, I am profoundly grateful to my advisor, Dr.John Stewart, for his invaluable guidance, insightful feedback, and unwavering commitment to my academic and professional growth. His mentorship has been instrumental in shaping the trajectory of my research. I would like to express gratitude to Dr.Gay Stewart. Her inspiring presence and motivating influence have significantly contributed to my accomplishments.

I extend my heartfelt thanks to my thesis committee members, Dr.Paul Miller, Dr. Weichao Tu, and Dr.Jake Follmer, for their expertise, constructive critiques, and valuable suggestions that have significantly enriched the quality of this thesis. A special thank you to the current members of our research group: Elaine Christman, John Hansen, Chris Wheatley, John Pace, Amanda Nemeth, Danielle Maldonado, and Brett Ballard, and the past members: Dr.Rachel Henderson and Dr.Cabot Zabriskie. The collaborative spirit and camaraderie within the group have made the research process both intellectually stimulating and enjoyable. I appreciate the dedication and support of the faculty, staff, and my fellow graduate students at the WVU Department of Physics and Astronomy. Their commitment to fostering a conducive academic environment has contributed significantly to my academic journey.

My heartfelt gratitude to my beloved husband, Chathu, whose unwavering support, patience, and encouragement have been my anchor throughout this challenging journey. I extend my gratitude to my mother, my father, and my family for their unconditional love, understanding, and support. Their sacrifices and belief in my abilities have been the driving force behind my pursuit of knowledge. I express my gratitude to my friend-sister, Sherry, for the ongoing love and kindness she consistently extends.

Last, but not least, my dear friends, who have provided a listening ear, and words of encouragement, and shared in both the triumphs and challenges—this friendship has been a source of joy and motivation.

In expressing my gratitude to each of you, I acknowledge that this achievement is the result of collective effort and support. Thank you for being an integral part of my academic and personal journey.

Contents

List of Tables

List of Figures

Chapter 1

Introduction to Physics Education Research

Undergraduate science and engineering education in the United States plays a crucial role in the development of undergraduates as professionals who are required to secure the next generation's future. Graduation rates among undergraduate students, regardless of their racial or ethnic backgrounds, are lower in the fields of science, technology, engineering, and mathematics (STEM) compared to other academic disciplines [3]. Making continuous improvements in undergraduate science and engineering education is the key to improving retention; institutions are actively striving to identify effective strategies to improve undergraduate education [4]. Physics Education Research (PER) studies issues in education specific to physics. PER applies a combination of qualitative and quantitative research methods, including surveys, interviews, classroom observations, and the analysis of student performance data. Researchers collaborate with physics educators to develop evidence-based teaching practices and educational materials that can improve the learning experience for students at all levels from K-12 to undergraduate and beyond.

In this chapter, we attempt to provide a brief review of the history of physics education research with a particular focus on conceptual understanding, assessment methods, and frequently used survey instruments for measuring both cognitive and non-cognitive aspects of physics learning.

1.1 History of Physics Education Research

PER emerged as a distinct and formal discipline to study teaching and learning physics in the late twentieth century; however, it had some notable contributors in the late nineteenth century. Pioneers like Ernst Mach [5] and Percy Bridgman [6] laid the groundwork for

later research in physics education. Physics professors such as Michael Faraday [7] and Heinrich Hertz [8] incorporated laboratory classes as an active learning method to teach students practical skills, reinforce theoretical concepts, and provide hands-on experience with scientific instruments which were difficult to achieve through traditional lectures. More recently, in the 1970s and 1980s, PER began to investigate student performance. Many researchers investigated a student's mathematical skill as a major factor in physics achievement [9–14]. Larkin and Brackett developed a math-review unit for the incoming undergraduate students [13]. In 1977, Hudson and McIntire showed that certain math skills were significantly correlated with a student's success in physics courses [10]. Although initial work focused on evaluating students' mathematical abilities, numerous studies have subsequently indicated that there are additional factors that play a significant role in influencing a student's success in physics. Champagne and Klopfer's model of students' achievement in a college physics course showed that only about 34% of the variance in students' mechanics achievement was explained by students' Newtonian physics understanding and math ability [14]. A large portion of the variance remained unexplained. Ibrahim and Hestenes showed that apart from their math skills, students' initial knowledge about physics also influenced their achievement in college physics [12]. Pallrand and Seeber concluded that the students' understanding of spatial reasoning influenced their achievement in introductory physics [15]. In the late 1970s, researchers started recognizing that students' cognitive abilities and their grasp of concepts emerged as key factors in achieving success in physics, complementing their mathematical skills [16, 17]. Champagne *et al.*'s study on the factors that influenced learning of classical mechanics contained an assessment tool they developed to evaluate these constructs, including students' preconceptions and logical reasoning [11]. The scores they collected from

their instrument called the Demonstration, Observation, and Explanation of Motion Test (D.O.E.) together with students' logical reasoning, measured using a 10-item questionnaire, were strongly correlated with the students' success in classical physics.

1.2 Modern Physics Education Research

This early research produced a model for Discipline-Based Education Research (DBER), in multiple fields in science and technology during the 1980s and 1990s. PER continued to grow as a research field [18, 19]. PER launched systematic research programs focused on student challenges at both the University of California, Berkeley, and the University of Washington [20].

In 1985, Halloun and Hestenes introduced the Mechanics Diagnostic Test to investigate the initial knowledge of college students [12]. This assessment aimed to evaluate their qualitative understanding of physics, covering essential concepts related to both kinematics and dynamics. An analysis of the test scores relative to students' performance in physics courses demonstrated the impact of students' initial beliefs about the natural world on their academic achievement. This pivotal development marked the beginning of research-based materials designed to assess students' conceptual understanding of physics. In 1992, Hestenes et al. used this prior research into student naive beliefs to develop the Force Concept Inventory(FCI) which tests students' understanding of kinematics and dynamics [21]. The FCI became the first widely used PER instrument and since, the most popular.

In 1998, Hake [22] showed that implementing Interactive Engagement (IE) methods in the classroom can significantly enhance the effectiveness of mechanics courses, surpassing the

outcomes achieved through traditional teaching methods. He conducted a survey analyzing pretest and post-test data, employing the FCI [21] across 62 introductory physics courses that collectively enrolled 6,542 students. To make comparisons among institutions with varying student demographics, Hake introduced the normalized gain metric $\langle g \rangle$, Eqn 1.1,

$$\langle g \rangle = \frac{\langle S_f \rangle - \langle S_i \rangle}{100 - \langle S_i \rangle}, \tag{1.1}$$

where $\langle S_i \rangle$ is the class pretest average and $\langle S_f \rangle$ is the class post-test average on a scale of 0 to 100.

Hake categorized a high-g as $\langle g \rangle \geq 0.7$, a medium-g as $0.7 > \langle g \rangle \geq 0.3$, and a low-g as $\langle g \rangle < 0.3$. Figure 1.1 displays a graph showing the average gain $\langle \text{gain} \rangle = \langle S_f \rangle - \langle S_i \rangle$ plotted against the pretest percentage scores for the 62 institutions examined in Hake's study. Figure 1.1 shows that classes implementing reformed instruction show enhanced normalized gains compared to those that do not. In this graph, the lines represent the thresholds for normalized gains, with steeper lines indicating larger improvements. This study played a pivotal role in driving the adoption of reformed instructional techniques, sometimes referred to as research-based instructional strategies (RBIS).

PER focuses on the study of how physics is taught and learned and seeks to understand the processes, methods, and outcomes of physics education to improve the effectiveness of teaching and learning in physics [19]. In their extensive synthesis of the field, Docktor and Mestre classified PER research into six distinct areas: conceptual understanding, problem-solving, curriculum and instruction, assessment, cognitive psychology, and attitudes and perceptions regarding learning and teaching [23].

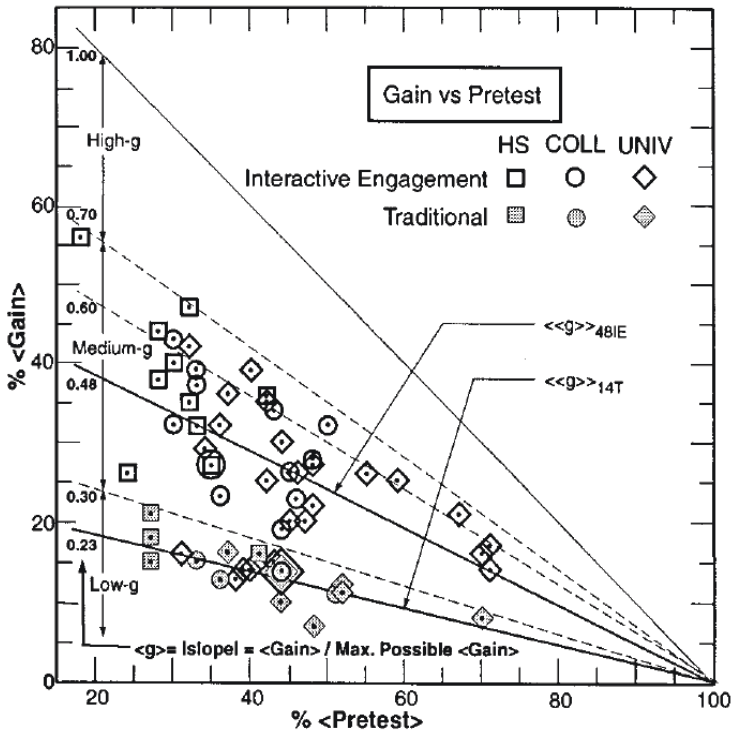


Figure 1.1: Traditional Instruction vs Interactive Engagement [22].

1.2.1 Problem-Solving

The study of problem-solving examines the process through which students employ their conceptual understanding, mathematical skills, logical reasoning, and various other abilities to solve physics problems [19]. This domain includes the strategies students employ when tackling physics problems [24, 25], how they transfer problem-solving practices from previous challenges to new ones [26–28], the types of representations they use during physics problem-solving and how these representations impact the process [29–32], distinctions between using mathematical skills in physics problem solving versus math problem solving [33–35], and the influence of diverse instructional strategies on students' problem-solving abilities [36, 37, 19].

1.2.2 Curriculum and Instruction

Research in curriculum and instruction focuses on the improvement of instructional methods and the development of curricular materials [19]. As research in this domain continues to advance, an abundance of materials have been created to enhance the active learning environment and foster interactions between instructors and students [38–43]. This work is influenced by advancements in communication and computer technology. This research includes upper-division undergraduate courses and graduate education, high school–university cross-curricular studies, courses for biology and pre-medical students, lab instruction, textbooks, and instructional technology [19].

1.2.3 Cognitive Psychology

This area explores the cognitive processes related to physics learning and physics problem-solving [19]. Jong *et al.* examined how novice problem solvers organize their knowledge and showed that good novice problem solvers arrange their knowledge around problem types [44]. Reif and Heller discussed how experts' problem-solving processes differ from those of novice students. In their study, they noted that experts swiftly reframe problems and outline solutions qualitatively before making mathematical decisions to compute numerical solutions. In contrast, novice students tend to assemble various mathematical formulas and initiate numerical calculations [45]. As intricate as cognition may be, in his study, Redish asserts that student knowledge systems exhibit a certain duality. They both appear disconnected and chaotic, while also rigid and robust. He also suggests that, to understand how students' knowledge systems work, emotional responses, motivations, and self-images

need to be taken into account [46].

1.2.4 Conceptual Understanding

One of the earliest and most extensively researched domains in PER examines students' conceptual knowledge [19]. As evidence accumulated that students have difficulty understanding some fundamental physics concepts, inquiries into the underlying causes of these challenges became common. These studies classified the challenges as misconceptions, naive conceptions, or alternative conceptions [47, 48]. Significant efforts have been dedicated to cataloging the preexisting notions that students possess before receiving formal physics instruction and distinguishing those that constitute misconceptions which conflict with correct scientific principles [48, 49]. While a substantial portion of these studies focus on topics within mechanics and kinematics, many studies have investigated electricity and magnetism, optics and light, thermal physics, and modern physics [50].

These misconceptions can subsequently hinder students' ability to grasp correct concepts in physics [51–53, 47]. Conceptual understanding research also focuses on designing, evaluating, and refining curricular interventions to address and rectify these misconceptions [54–57, 24]. This has led to substantial efforts to create, enhance, and validate survey instruments for measuring conceptual understanding.

1.2.5 Assessment

The development and evaluation of assessment tools are very important in PER, as they serve as a primary source of information for other research areas. The process of developing and validating these instruments involves a sequence of analyses initiated when researchers

recognize the need for an assessment tool. These analyses include qualitative studies related to the identified student difficulty, drafting and piloting the survey items, evaluating survey items and scores, comparing the scores with other measures over diverse populations, and finally exploring the score in complex statistical models beyond pretest/post-test models [19, 58, 59]. The progress in assessment research has resulted in the development of over 30 physics concept inventories for evaluating conceptual understanding. Furthermore, it has expanded its focus to gauge attitudes towards the learning of physics [12, 21, 60–62, 1].

As an example, the Force and Motion Conceptual Evaluation (FMCE) [62] is also a popular mechanics inventory that is used in a study in this thesis. The FMCE is a 43-item multiple-choice examination which [62] captures students' conceptual understanding of Newtonian mechanics. The FMCE contains items involving one-dimensional kinematics, Newton's three laws, and graphical reasoning. A decade after its introduction, the authors provided a modified scoring method consisting of a total score of 33, [63] which contained lesser survey items than the original inventory, removing some items and scoring some items as groups. We discuss how academic and non-cognitive factors affect college achievement using FMCE in Chapter 3 [64].

1.2.6 Attitudes and Perceptions Regarding Learning and Teaching

The final PER domain involves understanding student attitudes and perspectives about learning physics and how these attitudes relate to their performance in physics courses. This segment of PER focuses on research that investigates how attitudes and perceptions regarding learning and teaching evolve in response to instructional changes and interventions [19]. May and Etkina observed that students who experience significant conceptual improvements tend

to demonstrate more eloquent and epistemologically advanced reflections on their learning compared to students with limited conceptual gains. Three concept inventories, the FCI [65], the Mechanics Baseline Test [60], and the Conceptual Survey of Electricity and Magnetism [61] were used to measure the conceptual gains, and a series of weekly reports were used to measure their learning [66]. Perkins *et al.* observed that students' attitudes and beliefs towards physics learning substantially influence their retention in the course as well as in the discipline. Further, their beliefs positively correlated to their normalized conceptual learning gains [67]. Hammer investigated how students' epistemological beliefs could impact a high school teacher's perceptions of their students and instructional intentions. He also recommended gathering additional data through conceptual assessments and taking into account the distinct characteristics of each class being taught [68]. Several studies explore how faculty adopted the PER instructional strategies and materials, as well as the practical implementation of strategies such as Peer Instruction [69–73]. Various survey tools have been created to assess students' attitudes, beliefs, and perceptions regarding the learning and teaching of physics. These instruments continue to undergo refinement and enhancement to achieve greater precision in capturing specific facets of the learning experience [1, 74–76].

As an example, we discuss one of the most popular survey instruments to measure students' attitudes and beliefs about learning physics, the Colorado Learning Attitudes about Science Survey (CLASS). We will explore the CLASS in detail in Chapter 4. The CLASS, published in 2006, has become a widely used instrument for measuring students' beliefs about learning physics [1]. Student responses to CLASS items are compared to responses from an expert panel of physicists and scored as non-expert-like, neutral, or expert-like. Students enrolled in first-semester physics courses typically show attitudinal shifts away

from expert-like thinking [77]. Some pedagogical approaches including Physics by Inquiry [78], Peer Instruction [79], Physics in Everyday Thinking [80], and Modeling Instruction [81] have been shown to result in increases in expert-like responses.

Madsen *et al.* reported a tendency for physics majors to enter their undergraduate studies with more expert-like beliefs compared to non-majors and that these beliefs remain relatively stable throughout their undergraduate career. Chinese secondary students showed an overall decline in expert-like beliefs over a longer time scale [82]. Gray *et al.* [83] asked students to complete a modified CLASS that solicited students' attitudes and their beliefs about how physicists would respond to the same questions. Students' ideas about physicists' beliefs were quite stable over a semester, but their personal beliefs were most often negatively affected by instruction.

This chapter provided a brief general topical overview of PER. Additional topics will be discussed in more detail as they pertain to the research presented. The next chapters will provide a general overview of statistical methods (Chap. 2), explore the factors that influence pretest scores (Chap. 3), examine the CLASS in detail (Chap 4. to Chap. 6), and examine the relation of non-cognitive factors to achievement (Chap. 7).

Chapter 2

Statistical Methods

Mathematical methods, used to summarize, analyze, and interpret the observations, to support decision-making are statistics. Statistical methods are divided into two classes based on the purpose they serve: descriptive statistics, focusing on the characteristics of observed data, and inferential statistics, focusing on evaluating information to answer questions or make actionable decisions inferring the sample statistics of the population [84].

2.1 Descriptive Statistics

Organizing and summarizing information in a study to understand the characteristics of the sample are descriptive statistics. The complete set of individuals or items in a given group is the population and the characteristics that describe the population are population parameters. As it is generally impossible to include every individual or item of the population in a study, a selected set from the population is observed, the sample. The characteristics that describe the sample are sample statistics.

2.1.1 Scales of measurement

In a study, sample characteristics are represented using variables. The categories to which the values recorded for each variable belong are known as scales of measurement. There are four different scales of measurement used in this thesis:

1. Nominal scales identify something or someone (e.g. gender, student ID number, race). The numbers on the scale have no preferred order.

2. Ordinal scales convey comparisons, whether one value is less than, greater than, or equal to another. The numerical difference between the two values is not meaningful. For

example, the difference between the first and the second highest ranks of a Likert scale [85].

3. Interval scales are equidistant (intervals distributed in equal units) scales without a true zero. This indicates that there is no particular value that indicates the absence of the phenomenon. An example interval score would be the score on an IQ test.

4. Ratio scales are similar to interval scales as scores are distributed in equal units. Additionally, the distribution of the scores contains a true zero, which makes it an ideal scale in behavioral research as any mathematical operation can be performed on the scored values. Hours spent working on homework problems and the number of mid-semester exams taken are some examples of ratio scale variables used in education studies.

2.1.2 Measures of Central Tendency

To summarize a set of observations of a continuous variable (interval or ratio), a single value is selected to represent the distribution. The most common statistics used to characterize a distribution are measures of central tendency [84].

Mean

In this work, the sample mean (the sum of a set of scores in a distribution divided by the total number of scores) is presented as the measurement of the central tendency. Equation 2.1 calculates the sample mean M and Equation 2.2 the population mean μ [84].

$$M = \frac{\sum_i x_i}{n}, \tag{2.1}$$

$$\mu = \frac{\sum_i X_i}{N}, \tag{2.2}$$

where x_i are individual scores in the sample, X_i are individual scores in the population, n is the sample size, and N is the size of the population.

2.1.3 Variability

The mean provides a single measure to represent the distribution of the sample or the population; it does not capture how the scores differ from the mean or how the scores are distributed in the sample or the population. Variability is used to characterize the dispersion or the spread of scores in a distribution; the variability is captured by the variance, the average squared distance that individual scores deviate from the mean of the distribution. Population variance (σ^2) is calculated using equation 2.3.

$$\sigma^2 = \frac{\sum_{i=1}^{n}(X_i - \mu)^2}{N}, \tag{2.3}$$

where X_i is individual population scores, μ is the population mean, and N is the population size. Sample variance (s^2) is calculated using equation 2.4.

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - M)^2}{n-1}, \tag{2.4}$$

where x_i denotes individual sample scores, M is the sample mean, and n is the sample size. The numerator $\sum_{i=1}^{n}(x_i - M)^2$ is called the sum of squares (SS), the sum of the squared deviations of scores from the mean. To communicate the variability of each continuous

variable, the Standard Deviation (SD) is often presented with the sample mean as $M \pm SD$. The square root of Equation 2.3 is the population standard deviation (σ) and Equation 2.4 sample standard deviation (s) [84].

2.2 Inferential Statistics

Methods and techniques used to answer questions about sample attributes and their relationships in order to make actionable decisions and infer the population statistics from the sample are called inferential statistics. These procedures also allow the testing of hypotheses about the population [84].

2.2.1 Hypothesis testing

Hypothesis testing determines whether a hypothesis about population parameters is likely to be true [84]. To test a hypothesis the following four steps are performed:

Step 1 - State the hypothesis: First the null hypothesis H_0 is stated. The null hypothesis often states that there is no difference between two samples; $H_0 : \mu_1 = \mu_2$. Next an alternate hypothesis H_1 is stated that directly contradicts the null hypothesis, for example by stating the two means are not equal; H_1: $\mu_1 \neq \mu_2$.

Step 2 - Select the criteria for a decision: To set the criteria for a decision, the level of significance is defined, which is a criterion based on the probability of obtaining a statistic measured for the sample if the null hypothesis was true. This threshold probability for significance is denoted by (α). Traditionally, if the probability of obtaining a sample statistic is less than 5%, then the null hypothesis is rejected and the alternate hypothesis is retained.

Step 3 - Compute the test statistic: A test statistic with a known probability distribution based on the null hypothesis is then calculated. From this, the probability of obtaining the test statistic assuming the null hypothesis is true can be calculated. This value is the p value of the test.

Step 4 - Make a decision: Based on the test statistic calculated in Step 3, a decision is made either to retain the null hypothesis or to reject it. If the probability of observing the observed value is greater than the significance threshold, the null hypothesis is retained and the alternative hypothesis is rejected. If the probability of observing the value has a lower probability than the significance level, the null hypothesis is rejected and the alternative hypothesis is retained.

Two different errors are possible when testing a hypothesis: type I error (α) and type II error (β) [84]. Type I error occurs when the null hypothesis is rejected when it is true. The symbol α denotes the probability of type I error; α is equal to the level of significance. Researchers can directly control this error [84]. Type II error is associated with retaining a null hypothesis that is false. The symbol β denotes the probability of incorrectly retaining the null hypothesis [84]. The probability of rejecting the null hypothesis when it is false is called power in hypothesis testing [84]. This is denoted by $1 - \beta$, the opposite of type II error.

2.2.2 Effect Size

An effect for a single sample describes the difference between the sample mean and the population mean stated in the null hypothesis. If the test statistic fails to reject the null hypothesis, the effect is not statistically significant. To measure the practical difference

between the sample and the population, Cohen introduced the effect size [86]. Effect size measures the strength of a relationship between two variables in a sample or the degree of difference between two variables. Effect sizes are reported as small effects, medium effects, and large effects. A statistically significant effect is practically negligible when the effect size is smaller than a small effect [86, 84, 87].

There are many measures of effect sizes [88, 89, 87]. Cohen's d [86], a commonly used measure of effect size, measures effect size as the number of standard deviations the sample mean is shifted above or below the population mean stated in the null hypothesis as shown in Equation 2.5. Larger values indicated a larger shift or effect from the population mean. The effect size criteria for d: 0.2 is a small effect, 0.5 is a medium effect, and 0.8 is a large effect.

$$d = \frac{M - \mu}{\sigma}, \tag{2.5}$$

where M is the sample mean, μ is the population mean, and σ the population standard deviation.

2.2.3 t- test

The population standard deviation is often unknown. A t-test is a method used in hypothesis testing when the population standard deviation is not known. The t statistic is defined in Equation 2.6.

$$t = \frac{M - \mu}{\frac{s}{\sqrt{n}}} \tag{2.6}$$

where M is the sample mean, μ is the population mean, n is the sample size, and s is the sample standard deviation.

2.2.4 Correlation Analysis

Correlation analysis is a statistical method that examines the strength and direction of the relationship between two or more variables [90] assuming a linear relationship between the two variables. The magnitude of the correlation coefficient (r) represents how closely the data points are distributed around the best-fitting straight line. The correlation coefficient has a range of -1.0 and 1.0; the negative sign refers to inverse proportionality and the positive sign represents a direct proportionality. When the coefficient is zero, there is no linear relation present, and a scatter plot between the two variables is a random distribution of points. A correlation coefficient of magnitude 1.0 represents a perfect correlation where each data point falls on a straight line. The Pearson correlation coefficient is calculated as the ratio of the covariance of the two variables to the squared root of the product of each individual variance (Equation 2.7).

$$r = \frac{covariance\ of\ x\ and\ y}{variance\ of\ x\ and\ y\ separately} = \frac{\frac{\sum_{i=1}^{N}(x_i - M_x)(y_i - M_y)}{N-1}}{\sqrt{[\frac{\sum_{i=1}^{N}(x_i - M_x)^2}{N-1}\frac{\sum_{i=1}^{N}(y_i - M_y)^2}{N-1}]}} \tag{2.7}$$

To calculate the correlation between a continuous variable and a dichotomous variable, the point biserial correlation coefficient r_{pb} is used(Equation 2.8).

$$r_{pb} = \frac{(M_{y_1} - M_{y_2})\sqrt{(pq)}}{\sigma} \tag{2.8}$$

where M_{y_1} is the mean value of the continuous variable for level 1 of the dichotomous variable, M_{y_2} is the mean value of the continuous variable for level 2 of the dichotomous variable, p and $q = (1 - p)$ are the proportions of scores for each level of the dichotomous variable, and σ is the population standard deviation of the continuous variable.

2.2.5 Regression Analysis

This technique characterizes relations in data and the significance of the relations. Regression analysis attempts to describe the variance in the dependent variable using the variation in one or more independent variables [91]. Equation 2.9 represents a general multivariate regression equation.

$$y_i = \beta_0 + \beta_1 \cdot x_{1i} + \beta_2 \cdot x_{2i} + ... + \beta_n \cdot x_{ni} + \epsilon_i \tag{2.9}$$

where i indexes each independent measurement (usually each student), y_i is the dependent variable, x_{1i} through x_{ni} are the independent variables, ϵ_i is the error term of the regression equation; the variance of the dependent variable not explained by the independent variables, β_0 is the intercept of the regression line and β_1 through β_n are the slopes. The coefficients β_n are chosen the minimize the sum of the square errors, $\sum \epsilon_i$.

Significance testing is performed at two levels in regression analysis: (1) to test whether the regression equation predicts the variance of the dependent variable and (2) to test whether the relative contribution of each independent variable in explaining the variance of the dependent variable compared to the other independent variables. To test whether the regression equation predicts the variance of the dependent variable, an F test is used [92]. To test

if each independent variable significantly contributes to the prediction of the dependent variable compared to the other dependent variables, a t test is performed [93].

2.3 Factor Analysis

Factor Analysis, introduced by Spearman [94], is a statistical technique that analyses correlations among many observed variables and attempts to explain their variance using a smaller unobserved set of variables called latent variables [95]. This technique provides a more parsimonious representation of the observed variables [96]. Factors analysis can also be used to establish underlying relations between measured variables and latent constructs allowing the formation and refinement of theory and can provide construct validity evidence of self-reported scales [97]. When developing and validating an instrument, two types of factor analysis are utilized. Exploratory Factor Analysis (EFA) is used to deduce the factor structure by optimizing model fit and parsimony [98, 96, 95]. Confirmatory Factor Analysis (CFA) evaluates how well the factor structure fits the observations [95, 99]. In EFA, the factor structure is deduced from the data; in CFA, a theoretical model of the factor structure is compared with the data.

2.3.1 Exploratory Factor Analysis

EFA or unrestricted factor analysis is conducted at an early stage of a study to determine the dimensionality of an instrument when the researchers have an incomplete understanding of the constructs [96, 100]. Because it involves several statistical parameter estimations and the making of a sequence of decisions [100, 101], a five-step protocol is performed: suitability of data, extraction of factors, criteria to assist in determining factor

extraction, selection of rotational method, and interpretation and labeling was introduced by William *et al.* [97].

Suitability of data

As in any study, it is crucial to determine the minimum size of the sample. Out of several guiding rules of thumb [102–104] a general criteria of having at least 300 cases was suggested by Tabachnick [103]. Comrey and Lee suggested sample sizes of 100 as poor, 200 as fair, 300 as good, 500 as very good, and 1000 or more as excellent [104]. Several studies suggest determining the sample size based on the correlations among the items; stronger correlations with coefficients greater than 0.8 may require smaller samples [105, 106].

Determination of sample size also depends on the number of variables in the study, the "Sample to Variable Ratio" ($N : p$ ratio). This minimum ratio varies in different studies: $3 : 1$, $6 : 1$, $10 : 1$, $15 : 1$ and $20 : 1$ [97, 103, 101].

Because the number of variables directly affects the sample size, it is crucial to identify the number of variables with a significant influence. According to Tabachnick and Fidell [103], only variables with a correlation coefficient greater than 0.3 should be retained.

Extraction of factors

Numerous techniques to extract factors have been developed: principal factor analysis, principal axis factoring (PAF), and principal component analysis (PCA). PCA is the most common technique to extract factors when no a priori theory or model exists [97]. This study will apply PAF in order to replicate a previous study.

Criteria to assist in determining factor extraction

The purpose of factor extraction is to explain the responses to items in an instrument with fewer variables or factors; however, the optimal number of factors to describe a reasonable portion of the variance should be retained. Multiple methods have been developed to determine the optimal number of factors [107] and to confirm that number [97] including Kaiser's criteria ($eigenvalue > 1$), the scree test [108], the cumulative percent of variance extracted, and parallel analysis [109].

Selection of rotational method

The factor structure represents a reduced dimensionality basis for the data. Like other bases, it can be arbitrarily rotated. Rotation can clarify the relation between factors and the membership of factors. Two types of rotation are considered: orthogonal rotation when there are weak to no correlations among the factors and oblique rotation when correlations among the factors are considered or no a priori theory explains the correlations [97].

Interpretation and labelling

Once the factors are extracted, the researcher examines the items in each factor and identifies the construct measured by the factor. According to the attribute the items explain, a meaningful name is given to each factor [97].

To perform EFA, a set of linear equations relating the scores on n items to the scores on m postulated common factors ($m < n$) is introduced [110]. Each observed variable is expressed as a linear combination of m common factors and one error term. Let X_1, X_2,.....,X_n represent the standard scores on n observed items. $Y_1, Y_2, ..., Y_m$ represent

standard scores on m common factors. Let λ_{11}, λ_{12},....., λ_{nm} represent the the common factor loadings of the n observed variables on the m common factors. u_1, u_2,.....,u_n represent the residual error unique to each observed variable.

$$\begin{aligned} X_1 &= \lambda_{11}Y_1 + \lambda_{12}Y_2 + ...\lambda_{1m}Y_m + u_1 \\ X_2 &= \lambda_{21}Y_1 + \lambda_{22}Y_2 + ...\lambda_{2m}Y_m + u_2 \\ &... \\ X_n &= \lambda_{n1}Y_1 + \lambda_{n2}Y_2 + ...\lambda_{nm}Y_m + u_n \end{aligned} \tag{2.10}$$

Factor loadings are selected to minimize u_i for a given number of factors. EFA has several limitations including decision-making based on statistical criteria alone rather than using theoretical criteria [103, 101].

2.3.2 Confirmatory Factor Analysis

Researchers conduct Confirmatory Factor analysis (CFA) to support a proposed theory [99]. CFA partitions the variance of each item as a linear function of several factors. The variance of an item is separated into two segments: common variance, the variance of a set of factors common to other items, and the variance unique to the item.

Factor rotation is not applied in CFA because factor structure was established by the researcher based on theoretical considerations [111]. Because it provides explicit hypothesis testing, CFA is theoretically important.

2.4 Structural Equation Modeling

Structural Equation Modeling (SEM) models the relations among observed variables and latent variables in different theoretical models [112]. An observed variable is a variable we directly measure whereas a latent variable cannot be directly measured. A latent variable represents a hypothetical construct or an explanatory entity inferred from a collection of observed variables. SEM consists of a collection of statistical techniques to model relationships among a set of variables including continuous or discrete variables [113, 114]. SEM evaluates the fit of a theoretical model by comparing the variance-covariance matrices of a sample and that predicted by the theoretical model.

Structural equation models are often represented graphically using figures like that in Fig. 2.1. Directly measured variables are represented by rectangles and latent variables by ovals. Directed paths connect the observed and latent variables representing both multiple linear regression models and factor loadings. A single-head arrow represents a hypothesized directional causal effect and a two-headed curved dashed line represents a covariance.

To perform an SEM, the model is first specified based on a prior theoretical model. The model is then examined to determine if it is identified by testing the *order condition*, which states that a model is only identified when the number of free parameters in the model, calculated as $\frac{k(k+1)}{2}$ where k is the number of variables, is equal to or greater than the parameters estimated in the model. The free parameters in the model are the number of unique entries in the variance-covariance matrix. The number of parameters estimated is the combined number of path coefficients, error variances, independent variable variances, and correlations among the independent variables. When the number of free parameters equals

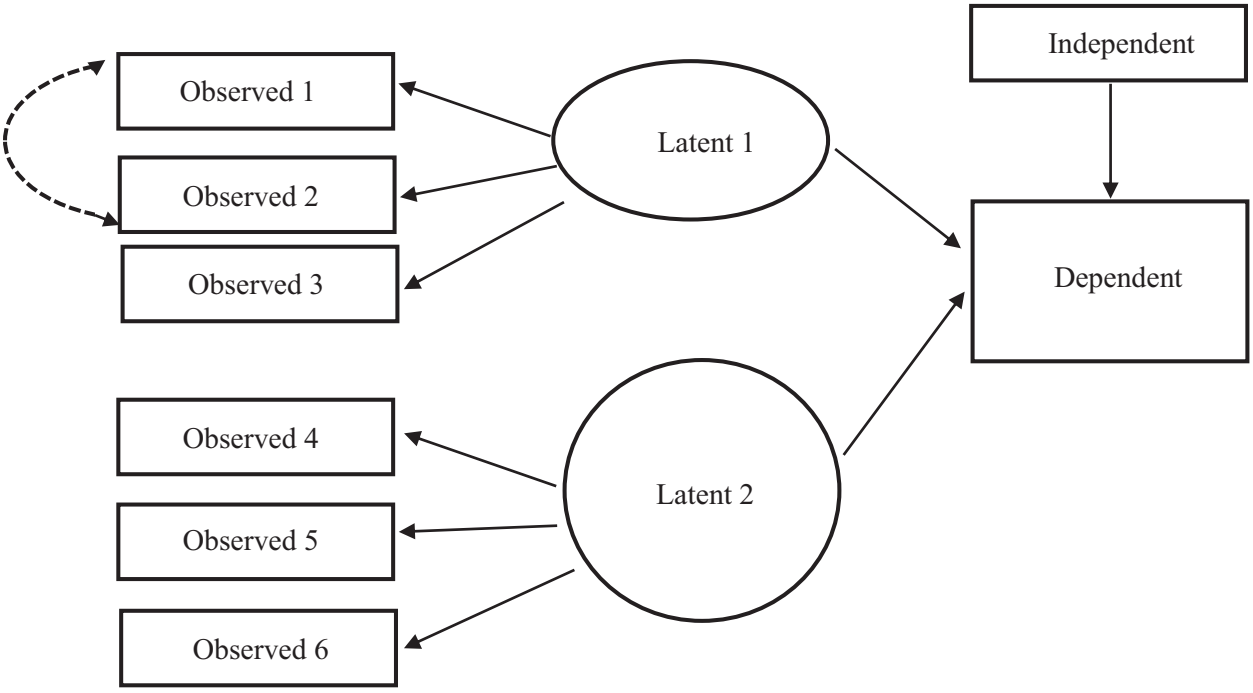


Figure 2.1: Structural Equation Model.

the number of parameters estimated, the model is considered just identified.

The parameters of the hypothesized model are then estimated using the maximum likelihood method and the model estimates are tested for statistical significance. The model fit is characterized calculating several fitness parameters including chi-squared test (χ^2) [115], Comparative Fit Index (CFI) [116], Tucker Lewis Index (TLI) [117], Root Mean Square Error of Approximation (RMSEA) [118] and Standardized Root Mean-square Residual (SRMR) [119].

Moderation and Mediation

Path models represent a subset of SEM models where all variables are directly observed (there are no latent variables). In multivariate regression, one independent variable can affect the relationship between another independent variable and the dependent variable.

When one independent variable affects the strength of the relationship between another independent variable and the dependent variable, such an independent variable moderates the relationship. To demonstrate moderation an interaction term is added to the regression [120] as shown in the Equation 2.11.

$$Dependent = \beta_0 + A \cdot (Independent) + B \cdot (Moderator) + C \cdot Independent \cdot Moderator + \epsilon \tag{2.11}$$

If coefficient C is statistically significant, then the moderation is significant. The moderator changes the slope of the independent variables from A to $A + C \cdot Moderator$.

Mediation is used to indicate when the effect of one or more independent variables is transmitted to the dependent variable through a different independent variable [121]. Baron and Kenny [120, 122] developed the framework to identify mediation in a regression relation. Mediated regression can be represented by a path model as shown in Fig. 2.2.

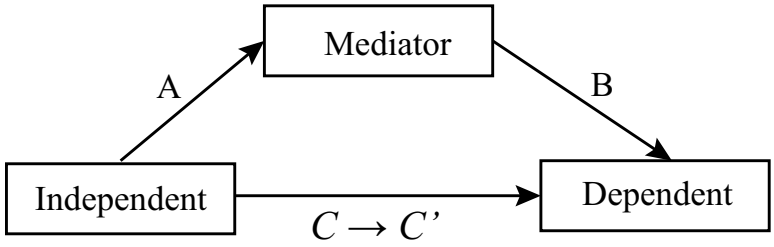


Figure 2.2: Path Model for Mediation.

The total effect C of the independent variable on the dependent variable is measured by Equation 2.12 where C is the relationship between the independent variable and the dependent variable without the presence of the mediator.

$$Dependent = \beta_0 + C \cdot (Independent) + \epsilon \tag{2.12}$$

The direct effect of the independent variable and the dependent variable is given by Equation 2.13

$$Dependent = \beta_0 + C' \cdot (Independent) + B \cdot Mediator + \epsilon \tag{2.13}$$

With the mediator present in the multivariate regression equation, the dependent variable acts through two paths, the direct effect, C' directly from the independent variable, and the indirect path, B through the mediator shown in Equation 2.14.

$$Mediator = \beta_0 + A \cdot (Independent) + \epsilon \tag{2.14}$$

The indirect path consists of two regressions, the independent variable explains the mediator, and the mediator explains the dependent variable. In Equation 2.14, A represents the first half of the indirect path, the independent variable explaining the mediator. A significant mediation exists, if the three regression coefficients, A, B, and C are statistically significant, and $C > C'$. Since Baron and Kenny, multiple refinements to mediation analysis have been proposed. Currently, best practice identifies significant mediation if bootstrapping creates a 95% confidence interval for the total indirect effect $(A \cdot B)$ which does not include zero.

This chapter reviewed some of the more common statistical methods applied in this work. More specialized methods will be discussed as they are introduced.

Chapter 3

Academic and Non-cognitive Factors Affecting College Achievement

*

*Parts of this chapter were published in "Hewagallage, D., Christman, E., & Stewart, J. (2022). *Examining the relation of high school preparation and college achievement to conceptual understanding.* Physical Review Physics Education Research, **18(1)**, 010149."

3.1 Introduction

This study was designed to explore the relationship between high school preparation, college achievement, and non-cognitive 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.

Comparing physics conceptual pretest and post-test scores to evaluate the development of conceptual understanding is prevalent in PER and is often used in physics classes [123, 124]. Often pretest and post-test scores are used as simple measures of physics conceptual knowledge [125]. However, recent studies suggest that treating the pretest and post-test scores as a simple measure of conceptual physics knowledge is incomplete because both measures show substantial relations with academic and non-academic factors. Physics conceptual pretest and post-test scores are related to students' general academic preparation measured by SAT and ACT scores [126–128]. Demographic factors such as gender, first-generation status, race/ethnicity, and the urban/rural status of the student's high school also are significantly related to pretest and post-test scores [126, 77, 129, 130]. Non-cognitive factors such as students' self-efficacy and personality are also related to conceptual pretest and post-test scores [131, 132].

A primary goal of PER is to identify potential supports and hindrances to the physics conceptual development to eliminate barriers and increase learning. To advance this goal, it is crucial to understand the initial physics conceptual understanding of the students [133, 128].

Pretest scores are often collected as a measurement of students' incoming physics preparation. The pretest is usually given in the first week of class before the material has been covered. Students often receive no prior notification to review and prepare for the pretest. These conditions could create the possibility that pretest scores do not capture the students' true prior preparation and may make non-cognitive factors such as self-efficacy, personality, and sense of belonging more important than in usual testing situations.

Using a conceptual pretest/post-test design to explore the effectiveness of an educational treatment has been common practice in the PER community for a long time [134]. Halloun and Hestenes, in 1985, used this methodology to show traditional instruction produced little additional conceptual understanding in college classes [12]. Their findings motivated the development of a catalog of student misconceptions about mechanics [135] that consequently resulted in the development of the Force Concept Inventory (FCI). The FCI is a 30-item multiple-choice instrument designed to measure students' conceptual understanding of Newtonian physics. Responses also present students with commonly selected incorrect answers [65]. The FCI pretest and post-test scores were used by Hake to show that the failure of traditional instruction to improve conceptual understanding was general [22]. The FCI has been used in many PER studies as a measurement of conceptual understanding [136–139] and, together with Hake's study, it provided a substantial impetus to the development of several other conceptual instruments including the Force and Motion Conceptual Evaluation (FMCE) [62], the Conceptual Evaluation of Electricity and Magnetism (CSEM) [61], and the Brief Electricity and Magnetism Assessment (BEMA) [140]. Current versions of many assessments are available at PhysPort [141].

3.2 Research Questions

This study seeks to answer the following research questions:

RQ1 What academic and non-cognitive factors are most important in predicting FMCE pretest scores?

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

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

3.3 Pretest as a control

Madsen *et al.* provided an extended overview of research-based assessment instruments in physics in 2017 [142]. The main purpose of using these instruments is to evaluate the efficacy of active teaching methods and other classroom interventions. The efficacy is often evaluated by applying a pretest followed by a post-test.

Multiple large studies have shown either the efficacy of reformed 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 [22]. 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 sci-

entific domains to show this result was general and not unique to physics classes [143]. 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 [144]. Many studies have reported gender differences in conceptual pretest and post-test scores; Madsen *et al.* provide a summary of this research [77].

3.3.1 Gain scores

Researchers in many different fields use pretest-post-test designs in their studies to understand the effectiveness of a treatment. This pretest–post-test design is analyzed in different ways to characterize the overall change [145]. The difference between the post-test score and pretest score, the actual gain, is often analyzed to understand the effect of a treatment [146]; however, in PER, the normalized gain, the ratio of actual gain to the maximum possible gain, is often reported. This statistic was popularized by the study by Hake comparing instructional methods [22]. 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 [147]. Either the actual gain, the normalized gain, or Cohen's d depend on the pretest score, the post-test score, and the relation of the pretest score to the post-test score. As such, all may be influenced by factors related to any of these quantities.

3.3.2 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 non-members of those groups including underrepresented minority students (URM), first-generation college stu-

dents (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 [126]. 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 and one the FCI. Stewart *et al.* partially replicated this work examining performance differences in FMCE post-test scores and course grades [127]. General high school preparation measured by ACT and SAT scores and prior preparation in physics strongly mediated demographic performance differences for First Generation College Students and Under Represented Minority students on both post-test scores and grades. No difference in course grades between men and women existed, so no mediation by course grade was possible. Gender differences in post-test scores were weakly mediated by ACT/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/SAT scores, and pretest scores in five large samples including two FMCE samples [129] 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 non-rural students on the FMCE pretest and post-test [130]. Pretest scores on the FCI, the FMCE, and the CSEM also correlate with post-instruction achievement measures (post-test score, test average, and course grades) differently for members of different demographic groups [148]. As such, general high school measures of achievement, ACT/SAT scores, and measures of prior physics knowledge explain some variation in a variety of physics achieve-

ment measures, but the variation explained is not consistent for different groups and much of the variation in the post-test performance of women is unexplained.

3.4 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 before 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 [126, 127].

Early work in PER predating the FCI investigated the effect of many cognitive factors on course grades or test averages including formal operational reasoning [149, 150], mathematics pretest scores [150, 151], and logical reasoning [151]. Meltzer showed the normalized gain on an electricity conceptual inventory was correlated with mathematics pretest and ACT/SAT mathematics percentile scores [128]. Coletta and Phillips found a positive correlation between Lawson's Classroom Test of Scientific Reasoning and FCI normalized gains [152]. 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

[152].

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 ($d = 0.19$) [153]. According to Hart and Cottle, math proficiency and high school physics background are vital for college achievement [154]. Hazari *et al.* investigated the relation of high school mathematics and sciences grades, taking AP calculus, instructional format, and some non-cognitive factors involving family support and found that many of these factors significantly predicted physics grades in college [155] 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 [156] finding both sets of variables as significant predictors of post-test scores. They also reported a 7% difference in FMCE post-test scores between students who had high school physics and students who did not; the effect of high school physics was larger for women, a 14% difference.

3.5 Factors affecting college achievement

Pretest and post-test scores measure a student's knowledge of physics. The research into the factors affecting the pretest score summarized above shows that they also measured 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.

3.5.1 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 [157] and with measures of general cognitive ability [158, 159]. The College Board touts the SAT's validity as a predictor of freshman-year GPA [160], while ACT has developed benchmarks for scores indicating a 50% chance of earning a B or higher in introductory college courses [161].

Although high school grades offer a less standardized measure of academic performance due to differing grading practices in different classrooms and schools [162], they are consistently stronger predictors of freshman-year GPA [163, 157], cumulative college GPA [164], and college completion [164–166] than SAT or ACT 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 SAT/ACT scores for college completion [166].

3.5.2 Non-cognitive factors

Many research studies have explored the influence of non-cognitive factors on college achievement including personality traits, motivational factors, and psychosocial contextual influence [167, 168].

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

[169] has been shown to affect students' performance and achievement in science classes [167, 170, 171]. A number of studies have found that male students have higher self-efficacy than female students in STEM classes [172–175]. Besterfield-Sacre *et al.* showed that these differences exist at the beginning of college using a study at 17 institutions [176]. 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 [177]. In physics, a study to explore the impact of Modeling Instruction on self-efficacy reported that traditional lecture classrooms negatively impact self-efficacy [178]. 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 in performance between men and women [179].

In this and many works, personality was characterized using the five-factor model with facets: agreeableness, conscientiousness, extraversion, neuroticism, and openness [180–182]. Personality has been shown to have a direct influence on academic performance and achievement [167, 168]. Stewart *et al.* reported that students' self-efficacy and personality were related to their college achievement [183]. 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 [167].

Beyond the non-cognitive 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 non-cognitive factors include mathematics anxiety [184, 185], science anxiety [186–188], stereotype threat [189], and attitudes toward science [77, Table I]. Theoretically, the non-cognitive 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.

3.6 Methods

3.6.1 The FMCE

The FMCE [62] measures conceptual understanding of Newtonian mechanics. The test consists of 43 multiple-choice items (excluding the energy items). After its introduction, Thornton *et al.* [63] 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.

Multiple studies have examined the item traits of the FMCE, including their factor structure [190, 130], network structure [191], and psychometric characteristics [190, 192, 193]. The assessed psychometric properties encompass reliability, item performance issues, and potential item bias. Additionally, more qualitative analyses have explored the instrument from the perspective of the resource framework [194].

Ramlo examined the factor structure and reliability of the FMCE using a sample of 146

students [190]. 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 [130]. 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 [192]. Many FMCE items had difficulty or discrimination within the range of problematic item functioning using CTT [195] on the pretest; fewer items were problematic on the post-test. Unlike the FCI, which contained many items unfair to women identified using DIF [196], the FMCE contained only one unfair item identified in both samples and this item was unfair to men.

3.7 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 [197].

The study was performed in the calculus-based introductory mechanics course taken by scientists and engineers. Student demographic and college performance measures were accessed from institutional records. Non-cognitive 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 during 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) or 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 $N = 1116$.

3.8 Instruments

Non-cognitive measures 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.

Personality

Personality was measured using the Big Five Inventory (BFI) which uses five facets to characterize personality: agreeableness, conscientiousness, extraversion, neuroticism, and openness [180–182]. 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 [198].

3.8.1 Self-Efficacy

Self-efficacy was measured using the Self-Efficacy for Learning and Performance subscale from the Motivated Strategies for Learning Questionnaire (MSLQ) [199]. The subscale

has strong validity [199] and is widely used [200]. This subscale asks the student to rate how much they agree with statements accessing self-efficacy in 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 [201].

3.8.2 Belonging

A student's sense of belonging in their physics class was assessed using three items adapted from Good, Rattan, and Dweck's "Math Sense of Belonging" instrument [202]. For example, students were asked how much they agreed 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 one could succeed on an examination.

3.8.3 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.

3.8.4 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 Advanced Placement (AP) and transfer classes was available for the institution studied. This was only available for AP or transfer (dual enrollment) classes that 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 scores on the AP test.

Students were asked about the first and second high school science classes taken including physics, chemistry, and biology. They were also asked to classify the level of each class as "Regular", "Honors", "AP", "Dual enrollment", and "Other advanced". Students were also asked to report the grades they received in each class. This generated a very complex set of data with multiple categories 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 classes were 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)".

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 classes 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.

3.9 Variables

Table 3.1 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.

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 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 the 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 AP physics test passed - grade B, C, D.
	HSP.APP.A	D	High school AP physics 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 the 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?

Table 3.1: List of Variables. Type indicates whether the variable is continuous (C) or dichotomous (D).

All calculations were performed with the "R" software system [203].

3.10 Results

3.10.1 Descriptive Statistics

Table 3.2 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 which 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). Analyses calculate changes against this variable. For a dichotomous variable in a panel, the correlation for a non-base 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

Panel	Variable	BL	%	M ± SD	r	p	R^2_{panel}	p_{panel}
	FMCE Pretest %			23.31 ± 18.3	1.00	0.000		
	FMCE Post-test %			46.61 ± 27.76	0.66	0.000		
Repeat	Is repeating physics class?		4.9		0.01	0.664	0.000	0.664
	College course completion %			94.01 ± 11.11	0.11	0.000		
	College GPA			3.35 ± 0.48	0.17	0.000		
College	College STEM classes taken			3.78 ± 0.91	-0.10	0.001	0.055	0.000
	College credit earned			27.11 ± 15.85	-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
	ACT or SAT Mathematics %			81.21 ± 13.93	0.27	0.000		
HS General	ACT or SAT Verbal %			75.32 ± 17.85	0.23	0.000	0.093	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 ± 3.22	0.12	0.012		
	Has transfer credit (not math or physics)		35.5		-0.02	0.438		
Transfer	Number transfer credits (not math or physics)			4.25 ± 4.22	-0.05	0.353	0.004	0.342
	Has transfer credit physics		1.9		-0.01	0.840		
	Has transfer credit calculus		9.7		-0.06	0.050		
	High school physics not taken	×	22.0		-0.22	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		
HS Physics	High school physics AP (test not passed) - B, C, D		11.4		0.35	0.000	0.175	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		
	High school last math grade A		58.4		0.08	0.007		
	High school last math not calculus	×	28.9		-0.15	0.000		
HS Math	High school last math calculus (not AP)		19.2		0.08	0.064	0.049	0.000
	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	Physics grade expectation A		41.3		0.24	0.000		
Expectation	Physics grade expectation B		41.0		0.12	0.002	0.044	0.000
	Physics grade expectation C, D, F, W	×	17.7		-0.15	0.000		
	Agreeableness			3.84 ± 0.57	-0.06	0.032		
	Conscientiousness			3.77 ± 0.55	-0.04	0.151		
Personality	Neuroticism			2.79 ± 0.76	-0.02	0.606	0.031	0.000
	Extraversion			3.19 ± 0.74	-0.12	0.000		
	Openness			3.65 ± 0.53	0.07	0.024		
	Gender (Female = 1)		29.1		-0.13	0.000		
Demographics	First-Generation (First-gen = 1)		15.9		-0.05	0.077	0.019	0.000
	URM (URM = 1)		7.0		-0.01	0.816		

Table 3.2: Descriptive Statistics. The base level of a set of dummy-coded variables is given by BL. 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 the pretest score and the probability that the correlation or a larger correlation occurred by chance p.

physics) by subsetting the data to only include students in these two groups. Other non-base variables in panels were handled similarly. The table also presents the R^2 values for a model regressing all variables in the panel on the pretest score as well as the significance of the

model. For panels with a single variable, $R^2_{panel} = r^2$. Panel regressions will be discussed further in Sec. 3.10.4.

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 ± 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 3.3: 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 are also presented. A t-test was performed to test 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.

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 [204]. Only a handful of variables have correlations with a medium to large effect; several variables in the HS Physics panel meet this criterion. 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 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 measures 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 3.3 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 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 a 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 3.3 provides support for the observations made about Table 3.2. 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 had a larger effect ($d = 0.77$).

The kind of high school physics taken has a dramatic effect on pretest scores with

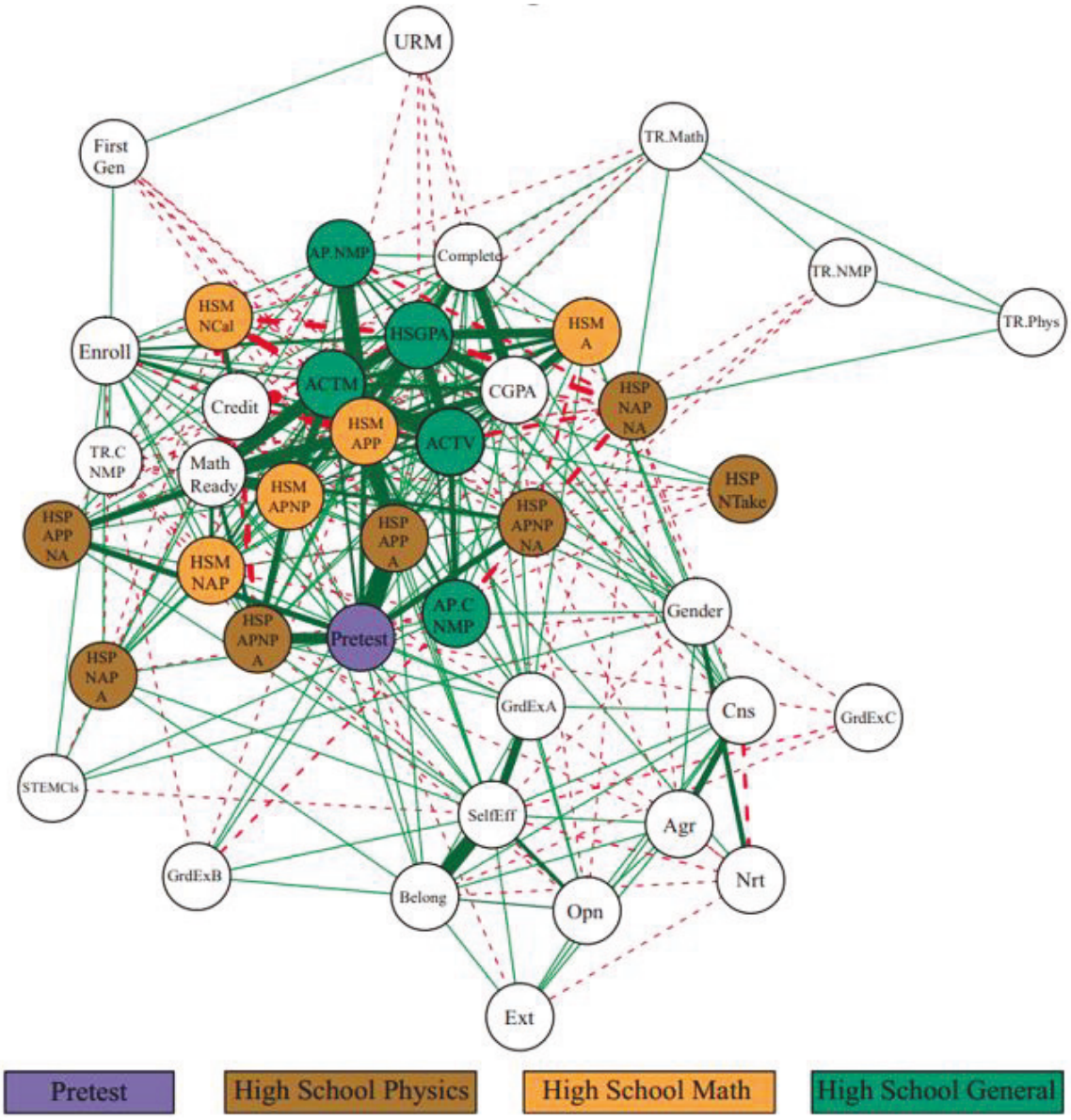
taking AP physics increasing pretest scores from $d = 0.79$ to an extraordinary $d = 2.7$ for students who passed the AP test and report earning an A in the AP class. Comparisons of the mean percent score for different methods of taking high school physics and different grade outcomes also show exceptional differences with students who do not take high school 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 3.2, 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.

3.10.2 Correlation Analysis

Figure 3.1 shows a visualization of the correlation matrix for variables in Table 3.1 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 represent a larger absolute value of the correlation. The visualization is rendered with the "qgraph" package in "R" that uses the force-direct graph visualization [205]. This representation is largely for visual effect, but it does allow the identification of groups of variables that are strongly inter-correlated. To produce the visualization, Hooke's law-like attractive forces are introduced between nodes with a spring constant proportional to the correlation coefficient. Repulsive Coulomb's law-like forces are introduced between all nodes with the same effective positive charge given to all nodes. The energy of the system is then minimized pushing weakly correlated nodes away from the system and drawing strongly correlated nodes together.

The correlation matrix helps highlight some general patterns in Tables 3.2 and 3.3. Pretest scores are most strongly correlated with taking AP high school physics for any grade

Figure 3.1: Correlation Matrix. Green (solid) lines represent positive correlations; red (dashed) lines negative correlations. Thicker lines represented stronger correlations.



as well as taking and passing AP calculus. In general, demographic characteristics, transfer credit and non-cognitive variables were weakly related to the pretest score. The non-cognitive variables do share some strong relations among themselves.

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

Table 3.4: Paneled variable importance. R_f^2 represents the variance explained when the variable is the only variable in the model. ΔR_l^2 represents the additional variance explained when the variable is the last variable added to the model. ΔR_s^2 is the average additional variance explained when adding the variable to a model subsampling the variable list to 5 variables. p_f in the p-value for the one-variable 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.

3.10.3 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 about the effect of multicollinearity 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 was. Measures of general high school academic success such as ACT 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, three measures of variable importance were calculated. The first uses the variable as the only independent variable in a linear regression predicting pretest score: R^2_f 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 the 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^2_l 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 that adds the variable of interest to the subset: the difference is captured by ΔR^2_s 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 3.1; the results are presented in Table 3.4.

There are a number of technical considerations involved in constructing this table. Some variables represent dummy variables coding a multi-level categorical variable such as the variables describing high school physics. For the base level of each group, the variable is handled normally. For variables other than the base level, the dataset is subsetted so only students in the base level or the level represented by the variable are included. The

dummy-coded categorical variables are not used to calculate the last-in or sampled variable importance.

Table 3.4 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 (R_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 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 1% additional variance when added last to the models.

3.10.4 Optimal pretest model

The full set of variables in Table 3.1 were used to predict the pretest score using multiple linear regression. The base level variables were removed because they are co-linear 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 Table 3.5. An optimal model, shown in Table 3.6,

was constructed by removing dependent variables that were not statistically significant at the $p < 0.05$ level. All variables in a panel were retained if one of the variables in the panel met this significance test. 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 the pretest score.

All regression tables present both the unstandardized regression coefficient B and 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 variables, 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 one 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.

The original model contained 36 independent variables, therefore, the 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

	B	SE	β	t	p
(Intercept)	-1.38	10.29	-0.64	-0.13	0.89373
Is repeating physics class?	9.95	2.41	0.54	4.13	0.00004
College course completion %	0.06	0.05	0.04	1.26	0.20663
College GPA	4.63	1.29	0.12	3.59	0.00034
College STEM classes taken	-1.06	0.54	-0.05	-1.94	0.05224
College credit earned	0.01	0.04	0.01	0.17	0.86112
College hours currently enrolled	-0.38	0.31	-0.04	-1.25	0.21071
Entered college math in calculus	3.02	1.27	0.17	2.38	0.01740
ACT or SAT Mathematics %	0.10	0.05	0.08	2.01	0.04493
ACT or SAT Verbal %	0.13	0.04	0.13	3.83	0.00013
High school GPA	-4.78	1.44	-0.11	-3.31	0.00096
Number AP classes (not math or physics)	-0.07	0.19	-0.01	-0.39	0.69730
Number transfer credits (not math or physics)	-0.05	0.15	-0.01	-0.36	0.71935
Has transfer credit physics	0.89	3.47	0.05	0.26	0.79810
Has transfer credit calculus	-1.62	1.68	-0.09	-0.97	0.33319
High school physics class not AP - B, C, D	5.29	1.59	0.29	3.33	0.00090
High school physics class not AP - A	4.56	1.34	0.25	3.41	0.00067
High school physics AP (test not passed) - B, C, D	10.25	1.78	0.56	5.75	0.00000
High school physics AP (test not passed) - A	15.57	1.69	0.85	9.20	0.00000
High school AP physics test passed - B, C, D	17.87	5.10	0.98	3.50	0.00048
High school AP physics test passed - A	30.02	3.04	1.64	9.87	0.00000
High school last math grade A	0.11	1.13	0.01	0.09	0.92587
High school last math calculus (not AP)	-0.40	1.45	-0.02	-0.28	0.78145
High school last math AP calculus (test not passed)	0.34	1.31	0.02	0.26	0.79562
High school last math AP calculus (test passed)	4.06	1.96	0.22	2.07	0.03827
Sense of belonging in physics	1.33	0.84	0.05	1.58	0.11519
Self-efficacy toward physics	1.01	0.88	0.04	1.14	0.25346
Physics grade expectation A	5.52	1.43	0.30	3.85	0.00012
Physics grade expectation B	2.49	1.35	0.14	1.85	0.06450
Agreeableness	-0.32	0.92	-0.01	-0.35	0.72616
Conscientiousness	-1.77	1.00	-0.05	-1.78	0.07597
Neuroticism	-0.27	0.72	-0.01	-0.37	0.71095
Extraversion	-2.51	0.68	-0.10	-3.71	0.00022
Openness	2.54	0.98	0.07	2.60	0.00932
Gender (Female = 1)	-4.25	1.16	-0.23	-3.67	0.00025
First-Generation (First-gen = 1)	0.48	1.31	0.03	0.37	0.71240
URM (URM = 1)	1.20	1.89	0.07	0.64	0.52514

Table 3.5: Full pretest model. B is the regression coefficient, SE is 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 $R^2 = 0.32$ [F(36, 1079) = 13.99, $p = 0.00000$] of the variance in pretest score.

presented in Table 3.7. 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 discussions.

	B	SE	β	t	p
(Intercept)	-0.46	6.92	-0.65	-0.07	0.94712
Is repeating physics class?	9.04	2.22	0.49	4.07	0.00005
College GPA	4.60	1.16	0.12	3.95	0.00008
College STEM classes taken	-1.31	0.52	-0.07	-2.51	0.01209
Entered college math in calculus	2.86	1.17	0.16	2.44	0.01490
ACT or SAT Mathematics %	0.10	0.05	0.08	2.13	0.03379
ACT or SAT Verbal %	0.14	0.03	0.13	4.00	0.00007
High school GPA	-5.23	1.37	-0.13	-3.83	0.00014
High school physics class not AP - B, C, D	5.39	1.55	0.29	3.47	0.00054
High school physics class not AP - A	4.83	1.31	0.26	3.68	0.00024
High school physics AP (test not passed) - B, C, D	10.40	1.75	0.57	5.95	0.00000
High school physics AP (test not passed) - A	15.98	1.66	0.87	9.64	0.00000
High school AP physics test passed - B, C, D	17.91	5.03	0.98	3.56	0.00039
High school AP physics test passed - A	30.73	2.99	1.68	10.28	0.00000
High school last math calculus (not AP)	-0.55	1.42	-0.03	-0.39	0.69929
High school last math AP calculus (test not passed)	0.41	1.28	0.02	0.32	0.74808
High school last math AP calculus (test passed)	4.22	1.88	0.23	2.24	0.02507
Physics grade expectation A	6.28	1.36	0.34	4.60	0.00000
Physics grade expectation B	2.89	1.33	0.16	2.18	0.02981
Extraversion	-2.39	0.65	-0.10	-3.70	0.00023
Openness	2.95	0.91	0.09	3.23	0.00128
Gender (Female = 1)	-4.99	1.06	-0.27	-4.69	0.00000

Table 3.6: Optimal pretest model removing non-significant independent variables. B is the regression coefficient, SE is 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 $R^2 = 0.31$ [F(21, 1094) = 23.29, $p = 0.00000$] of the variance in pretest score.

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

Table 3.7: Optimal pretest model with Bonferroni correction. B is the regression coefficient, SE is 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.

For the dichotomous variables, repeating the class is near a medium effect as is 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.

3.10.5 Optimal post-test model

The same set of variables including the pretest score was used to predict the post-test score. The full regression model is presented in Table 3.8. An optimal model was constructed by removing independent variables that were not significant at the $p < 0.05$ level and is presented in Table 3.9. This model explained $R^2 = 0.56$ of the variance in the post-test score. This model was not statistically different from the full model using an ANOVA test [F(21, 1078) = 1.39, $p = 0.1103$]. As before, a Bonferroni-corrected model was constructed, which was statistically inferior to the optimal model, [F(9, 1099) = 5.42, $p = 0.00000$], but explained 54% of the variance. We will focus on this model presented in Table 3.10 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 no additional effects of high school physics on the post-test score. To further test this conclusion, the post-test models were fit without using the pretest score as an independent variable. The model with Bonferroni correction is shown in Table 3.11. Without the pretest score in the model, high school physics is a strong predictor of post-test

	B	SE	β	t	p
(Intercept)	9.35	12.42	0.06	0.75	0.45179
FMCE Pretest %	0.79	0.04	0.52	21.45	0.00000
Is repeating physics class?	-4.09	2.93	-0.15	-1.40	0.16272
College course completion %	0.00	0.06	0.00	0.04	0.97091
College GPA	3.74	1.56	0.06	2.39	0.01697
College STEM classes taken	-1.44	0.66	-0.05	-2.19	0.02881
College credit earned	-0.04	0.05	-0.02	-0.72	0.47214
College hours currently enrolled	-0.59	0.37	-0.04	-1.59	0.11296
Entered college math in calculus	2.15	1.53	0.08	1.40	0.16066
ACT or SAT Mathematics %	0.20	0.06	0.10	3.34	0.00088
ACT or SAT Verbal %	0.22	0.04	0.14	5.07	0.00000
High school GPA	-4.85	1.75	-0.08	-2.77	0.00576
Number AP classes (not math or physics)	0.39	0.23	0.04	1.72	0.08585
Number transfer credits (not math or physics)	0.19	0.18	0.02	1.03	0.30119
Has transfer credit physics	4.92	4.19	0.18	1.17	0.24106
Has transfer credit calculus	-2.53	2.02	-0.09	-1.25	0.21221
High school physics class not AP - B, C, D	-0.07	1.93	-0.00	-0.04	0.97137
High school physics class not AP - A	2.03	1.62	0.07	1.25	0.21203
High school physics AP (test not passed) - B, C, D	3.58	2.18	0.13	1.64	0.10153
High school physics AP (test not passed) - A	0.86	2.12	0.03	0.41	0.68482
High school AP physics test passed - B, C, D	2.03	6.19	0.07	0.33	0.74334
High school AP physics test passed - A	2.78	3.83	0.10	0.72	0.46898
High school last math grade A	3.41	1.36	0.12	2.50	0.01253
High school last math calculus (not AP)	1.16	1.75	0.04	0.66	0.51024
High school last math AP calculus (test not passed)	0.26	1.58	0.01	0.16	0.86996
High school last math AP calculus (test passed)	5.21	2.36	0.19	2.20	0.02788
Sense of belonging in physics	0.23	1.02	0.01	0.23	0.81858
Self-efficacy toward physics	2.74	1.07	0.07	2.57	0.01032
Physics grade expectation A	-5.05	1.74	-0.18	-2.90	0.00382
Physics grade expectation B	-4.83	1.63	-0.17	-2.97	0.00306
Agreeableness	0.81	1.11	0.02	0.73	0.46269
Conscientiousness	-1.76	1.21	-0.03	-1.46	0.14558
Neuroticism	-0.27	0.87	-0.01	-0.31	0.75654
Extraversion	-2.33	0.82	-0.06	-2.83	0.00467
Openness	2.42	1.18	0.05	2.05	0.04061
Gender (Female = 1)	-10.35	1.41	-0.37	-7.35	0.00000
First-Generation (First-gen = 1)	3.25	1.59	0.12	2.05	0.04073
URM (URM = 1)	-4.70	2.28	-0.17	-2.06	0.03948

Table 3.8: Full post-test model. B is the regression coefficient, SE is 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 $R^2 = 0.57$ [F(37, 1078) = 38.5, $p = 0.00000$] of the variance in post-test score.

scores. 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.

	B	SE	β	t	p
(Intercept)	0.53	8.11	0.17	0.07	0.94783
FMCE Pretest %	0.81	0.03	0.53	23.89	0.00000
Is repeating physics class?	-5.84	2.68	-0.21	-2.18	0.02966
College GPA	2.83	1.42	0.05	1.99	0.04713
College STEM classes taken	-1.61	0.63	-0.05	-2.56	0.01054
ACT or SAT Mathematics %	0.24	0.06	0.12	4.28	0.00002
ACT or SAT Verbal %	0.24	0.04	0.16	5.99	0.00000
High school GPA	-4.17	1.69	-0.07	-2.48	0.01347
High school last math grade A	3.38	1.32	0.12	2.55	0.01081
High school last math calculus (not AP)	0.77	1.70	0.03	0.45	0.64950
High school last math AP calculus (test not passed)	0.58	1.51	0.02	0.39	0.69957
High school last math AP calculus (test passed)	6.50	2.22	0.23	2.93	0.00343
Self-efficacy toward physics	3.36	0.87	0.09	3.84	0.00013
Physics grade expectation A	-5.52	1.73	-0.20	-3.19	0.00146
Physics grade expectation B	-5.21	1.62	-0.19	-3.21	0.00134
Extraversion	-2.05	0.77	-0.05	-2.65	0.00805
Gender (Female = 1)	-10.55	1.31	-0.38	-8.03	0.00000

Table 3.9: Optimal post-test model removing non-significant independent variables. B is the regression coefficient, SE is 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 $R^2 = 0.56$ [F(16, 1099) = 86.55, $p = 0.00000$] 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
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
GrdExA	-5.32	1.73	-0.19	-3.07	0.00217
GrdExB	-5.78	1.64	-0.21	-3.53	0.00044
Gender	-10.13	1.30	-0.37	-7.79	0.00000

Table 3.10: Optimal post-test model with Bonferroni correction. B is the regression coefficient, SE is 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 $R^2 = 0.54$ [F(7, 1108) = 184.25, $p = 0.00000$] of the variance in post-test score.

The Bonferroni corrected optimal model (Table 3.10) 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 the pretest score, gender, and ACT or SAT scores explains 53% of the variance and is shown in Table 3.12. One can then progressively remove variables to determine how much variance each explains. Removing gender produced

	B	SE	β	t	p
(Intercept)	-13.21	9.17	-0.12	-1.44	0.14994
College GPA	6.92	1.65	0.12	4.20	0.00003
ACT or SAT Mathematics %	0.38	0.07	0.19	5.80	0.00000
ACT or SAT Verbal %	0.35	0.05	0.22	7.07	0.00000
High school GPA	-8.20	2.00	-0.13	-4.10	0.00004
High school physics class not AP - B, C, D	2.51	2.26	0.09	1.11	0.26639
High school physics class not AP - A	5.86	1.90	0.21	3.08	0.00209
High school physics AP (test not passed) - B, C, D	10.93	2.51	0.39	4.35	0.00001
High school physics AP (test not passed) - A	13.17	2.40	0.47	5.49	0.00000
High school AP physics test passed - B, C, D	14.85	7.36	0.53	2.02	0.04382
High school AP physics test passed - A	25.96	4.39	0.93	5.92	0.00000
High school last math calculus (not AP)	-0.08	2.06	-0.00	-0.04	0.97039
High school last math AP calculus (test not passed)	0.37	1.83	0.01	0.20	0.83906
High school last math AP calculus (test passed)	9.79	2.72	0.35	3.61	0.00033
Self-efficacy toward physics	5.34	1.00	0.14	5.34	0.00000
Extraversion	-4.15	0.93	-0.11	-4.47	0.00001
Gender (Female = 1)	-14.10	1.57	-0.51	-9.01	0.00000

Table 3.11: Optimal post-test model without pretest scores with Bonferroni correction. B is the regression coefficient, SE is 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 $R^2 = 0.35$ [F(16, 1099) = 37.59, $p = 0.00000$] of the variance in post-test score.

a model that explained 50% of the variance (Table 3.13); gender explained 3% additional variance controlling for pretest score and ACT scores. Removing ACT or SAT mathematics and verbal percentile scores produced a model that explained 44% of the variance using only the pretest score. Pretest alone explained 44% of the variance in post-test score; ACT scores explained an additional 6% of the variance controlling for pretest score. This should not apply because these variables are independent; a model with ACT or SAT scores but not pretest scores explains 18% of the variance in post-test scores.

	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 3.12: Post-test model with pretest, gender, and ACT or SAT. B is the regression coefficient, SE is 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 $R^2 = 0.53$ [F(4, 1111) = 310.03, $p = 0.00000$] of the variance in 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

Table 3.13: Post-test model with pretest and ACT or SAT. B is the regression coefficient, SE is 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 $R^2 = 0.50$ [F(3, 1112) = 368.56, $p = 0.00000$] of the variance in post-test score.

3.11 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 non-cognitive 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 scores. High school physics has been investigated in other works with varying outcomes [154, 153, 155, 156]. 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 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 the pretest score. The difference in pretest scores 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 [153]. 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. Having AP physics taken, AP test passed,

and an A earned increased the pretest score 37% 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 captures both the kind of high school physics taken and the student's grade explain 17.5% of the variance in the 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 scores above the threshold of a small effect as shown in Table 3.2. Beyond high school physics, ACT or 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 scores between the two levels of the variable (Table 3.3). 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. 3.1.

Measures of variable importance, Table 3.4, 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 3.7, 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 had small effects as did CGPA and ACT or SAT verbal score.

RQ2: What academic and non-cognitive 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 3.10, which contained the pretest score as a variable did not contain any high school physics variables; the 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. Post-test scores also depended on general high school academic preparation measured by ACT or 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/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/SAT scores. As such, the majority of the variance in post-test scores is explained by pretest scores, but some prior academic achievement variables and demographic variables are also important.

RQ3: Do academic and non-cognitive 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 non-cognitive factors. The overall gender difference in the pretest score was 5.1%; the difference grew to 13.8% on the post-test. The models controlling for both high school preparation and non-cognitive factors (Table 3.7 and Table 3.10) failed to account for the gender difference. If non-cognitive factors 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 3.7 showed a gender difference of 5.6% correcting for all these factors. The gender regression 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 differences in pretest scores was explained by non-cognitive 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 non-cognitive 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 *et al.*'s [126] finding that prior preparation variables fully mediated gender differences in final examination scores. It is consistent with Stewart *et al.*'s observation that much of the gender difference in post-test scores is unexplained by the same factors [127].

3.12 Implications

This study examined the features predicting pretest scores with a large sample and an extensive set of high school level, college level, and non-cognitive 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, uncertainty in pretest scores will generate uncertainty in the absolute gain and the normalized gain.

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 were 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 3.7, 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 prior knowledge of

physics.

For the optimal post-test regression model, Table 3.10, the majority of the variance was explained by pretest scores (44%), ACT/SAT scores explained an additional 6% of the variance, and gender 3%. All other variables only explained 1% together. As such, the pretest score captures most of the effect of prior preparation and non-cognitive effects on post-test scores 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 [147]; specifically how can and should the normalized gain popularized by Hake [22] be updated? Nissen *et al.* showed the normalized gain was biased toward students with higher pretest scores [147]. This work showed that both pretest and post-test scores were related to general high school level achievement measured by ACT/SAT scores and that pretest scores were 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 [126, 127].

Examining the differences in pretest scores by level of high school preparation in Table 3.3 shows that there is a broad spectrum of prior preparation in physics in the class studied. It is important to consider this when designing activities in the class and interpreting the results of the 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 curricula such as the International Baccalaureate (IB) program could not produce similar results. There were insufficient students in these programs 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.

3.13 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 to determine if the results are general.

3.14 Conclusion

This study applied correlation analysis and linear regression analysis to understand the relation of high school preparation, college achievement, and non-cognitive 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 non-cognitive 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 the 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 expectations, and self-efficacy were significant in predicting post-test scores 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 factors on post-test scores. Gender differences observed in FMCE pretest and post-test scores were changed little by controlling for either high school preparation or non-cognitive factors.

Chapter 4

Introduction to the Colorado Learning Attitudes about Science Survey

4.1 Introduction

The Colorado Learning Attitudes about Science Survey (CLASS) is a popular survey instrument used to gauge students' attitudes and beliefs about physics and the learning of physics. The CLASS consists of 42 statements with which a student may strongly agree, agree, remain neutral, disagree, or strongly disagree, creating a five-point Likert scale [85]. A panel of experts rated whether high scores or low scores on each item represented expert-like beliefs producing a 3-level scale: non-expert-like, neutral, and expert-like. The authors of the CLASS reported that exploratory factor analysis suggested eight factors, each consisting of four to eight items [1]. The factors identified were Real World Connection, Personal Interest, Sense Making/Effort, Conceptual Connections, Applied Conceptual Understanding, Problem-Solving General, Problem Solving Confidence, and Problem-Solving Sophistication as shown in Table 4.1. Twenty-seven of the items loaded onto a factor; 16 of these 27 items loaded onto two or more factors. Nine items did not load onto a factor and six additional items were not scored. The overall instrument and each subscale (factor) are then given a score representing the percentage of expert-like responses (% favorable) [1].

Category	Survey Items
Real World Connection	28,30,35,37
Personal Interest	3,11,14,25,28,30
Sense Making/Effort	11,23,24,32,36,39,42
Conceptual Connections	1,5,6,13,21,32
Applied Conceptual Understanding	1,5,6,8,21,22,40
Problem Solving General	13,15,16,25,26,34,40,42
Problem Solving Confidence	15,16,34,40
Problem Solving Sophistication	5,21,22,25,34,40
Not Scored	4,7,9,31,33,41

Table 4.1: Subscales introduced in the original CLASS [1].

Example items from each subscale follow:

- Real World Connection: "Learning physics changes my ideas about how the world works." and "Reasoning skills used to understand physics can be helpful to me in my everyday life."
- Personal Interest: "I think about the physics I experience in everyday life." and "I am not satisfied until I understand why something works the way it does."
- Sense Making/Effort: "I am not satisfied until I understand why something works the way it does.", and "In doing a physics problem if my calculation gives a result very different from what I'd expect, I'd trust the calculation rather than going back through the problem."
- Conceptual Connections: "A significant problem in learning physics is being able to memorize all the information I need to know." and "After I study a topic in physics and feel that I understand it, I have difficulty solving problems on the same topic."
- Applied Conceptual Understanding: "When I solve a physics problem, I locate an equation that uses the variables given in the problem and plug in the values."
- Problem-Solving General: "If I get stuck on a physics problem on my first try, I usually try to figure out a different way that works." and "Nearly everyone is capable of understanding physics if they work at it."
- Problem-Solving Confidence: "I can usually figure out a way to solve physics problems." and "If I get stuck on a physics problem, there is no chance I'll figure it out on my own."

- Problem-Solving Sophistication: "If I don't remember a particular equation needed to solve a problem on an exam, there's nothing much I can do (legally!) to come up with it." and "If I want to apply a method used for solving one physics problem to another problem, the problems must involve very similar situations."

Item 31 was used to identify the good faith effort by asking students to select a specific response; it was not included in subsequent analysis. Five items (4, 7, 9, 33, 41) could not be classified as expert-like or non-expert-like; they did not have an "expert" response or are statements that are not useful in their current form. These items were not scored in Adams *et al.*.

A version of the CLASS with minor modifications has also been used in chemistry and biology classes for similar purposes [206, 207]. Another survey instrument, the Colorado Learning Attitudes about Science Survey for Experimental Physics (E-CLASS), was developed to measure the attitudes toward experimental physics in undergraduate physics laboratories [207].

The CLASS has been used in a broad variety of studies. Kost *et al.* used CLASS scores to explore gender differences in conceptual post-test scores [156]. Ding found a significant causal relationship between pretest scores on the CLASS and normalized gains on a mechanics conceptual inventory [208]. Traxler and Brewe disaggregated CLASS data to examine the effects of Modeling Instruction on the attitudes of women and underrepresented minorities [209]. Baily and Finkelstein used the CLASS to monitor students' development of a probabilistic quantum-mechanical perspective in response to modern physics instruction [210]. Gire *et al.* used the CLASS to compare the views of introductory physics students

majoring in physics and engineering [211]. The physics CLASS has been used as a basis for the development of similar surveys for chemistry [206], biology [207], laboratory practices [212], and informal science education experiences [213]. Sawtelle *et al.* found that students at a predominately Hispanic university overwhelmingly interpreted the statements as intended and flagged one item, item 21, as misinterpreted by over one-third of students [214].

In 2014, Douglas *et al.* attempted to replicate the factor structure published with this instrument [2]; they were unsuccessful and reported that a three-factor model optimized fit statistics. Even this reduced model did not meet modern criteria for a well-fitting model. A modified instrument was proposed containing only the 15 items organized into three subscales [2]. This factor structure is shown in Table 4.2. Kontro and Buschhüter attempted confirmatory factor analysis and found that some modifications were needed to obtain acceptable fit statistics; they ultimately developed a 14-item instrument, discarding an item that loaded on two factors and including the *post hoc* adjustments proposed by Douglas *et al.* [215].

Category	Survey Items
Personal Application and Relation to Real World	3,14,25,28,30,37
Problem Solving/Learning	5,21,22,34,40
Effort/Sense-Making	23,24,29,32

Table 4.2: Subscales identified in Douglas *et al.'s* refactoring of the CLASS [2].

4.2 Research questions

The present study attempts to replicate and then extend the analysis of Douglas *et al.*. The following chapters will discuss two research questions.

RQ1: Can the Douglas *et al.* or the Adams *et al.* factor structure be replicated? How can the factor model be improved to produce a well-fitting model?

RQ2: What is the optimal factor model for the CLASS instrument?

Chapter 5 will address RQ1 and Chapter 6 will address RQ2.

Chapter 5

Replicating Douglas *et al.*'s factor analysis of the CLASS

5.1 Introduction

Multiple studies have suggested that the factor structure originally published for the Colorado Learning Attitudes about Science Survey (CLASS) is not a good fit for the instrument. This chapter provides a replication of the analysis of Douglas *et al.* which found a 3-factor solution for the instrument. The first part of this chapter describes the method used in the Douglas *et al.* study and the latter part presents our attempt to replicate the results of the study.

5.1.1 The Douglas Procedure

Douglas *et al.* followed a detailed procedure, which we call the "Douglas procedure" to extract an optimal factor structure for the instrument. The step-by-step procedure follows.

1. The dataset was split into two approximately equally sized sets: the exploratory dataset and the confirmatory dataset. The exploratory dataset was used for steps 2 to 7. The confirmatory dataset was used for steps 8 and 9.

2. Bivariate correlations between each pair of items were calculated. The bivariate correlation is the correlation between the scores of the two items. Items with a maximum bivariate correlation of less than 0.275 were removed.

3. Item-total correlation was computed for each item. This is the correlation between the average score on all items to the score of a single item. Items with item-total correlation below 0.20 were removed.

4. EFA was performed and the number of factors was identified (we presume this hap-

pened at this point because it was required by the next step). For our replication, the scree plots were used to identify the number of factors. The Douglas paper is not clear on the method used. Principal axis factoring was used to extract the factors along with Promax rotation because the factors were not orthogonal [216].

5. The communality coefficient, or the shared variance of each item, was calculated. The communality coefficient is the sum of the squares of the factor loadings of the item. Items with a communality coefficient less than 0.30 were removed iteratively. After each item removal, the EFA was performed again, and item communalities were rechecked.

6. Items with factor loadings less than 0.30 on all factors were removed.

7. Cronbach's alpha was calculated for each factor to characterize its internal reliability.

8. Confirmatory Factor Analysis was then performed on the factor structure using the reserved dataset reserved for CFA. Model fit statistics were compared to standards for acceptable fit.

9. If the model did not meet standards for an acceptable fix, modification indices were used to adjust loadings and covariances to improve model fit.

There were multiple points where the paper was vague about what was done; we had to make educated guesses at these points. There were also a few minor points where we disagreed with the process. We assume that the full five-point Likert scale was used for each item. We further assume, because no negative loadings were reported, that reverse-coded items (where the high values on the scale were in-expert-like) have been recorded to reverse the scale. Douglas included all items; we removed the items not scored in the original CLASS

paper. Douglas removed low bivariate or item-total correlation items from the confirmatory dataset; we feel it is bad practice the modify the confirmatory dataset. We, instead, used a bootstrapping method on the exploratory dataset to identify items robustly below the threshold. Douglas used communality coefficients which require a factor model but does not discuss how the number of factors was identified or whether the number of factors was reconsidered after each item was removed for low commonality. We assume the scree plot was used to identify the number of factors and that this was done once with the full dataset. We also chose to examine post-instruction records. Previous analysis of this dataset showed little change in CLASS scores pre- and post-instruction. Note, that item 31 was a marker question designed to identify students who were not answering seriously and was excluded from the analysis.

5.2 Summary of the Douglas Results

5.2.1 Sample

The sample Douglas *et al.* used in their study contained 3,844 college students enrolled in an introductory-level calculus-based physics course at a large midwestern university in the US. Data were collected pre-instruction over eight different semesters starting with the Spring 2006 semester. Pre-instruction data were used to avoid any effect of the instruction on the structure extracted for the instrument. The majority of the students were freshmen and sophomore engineering majors. The CLASS survey was made available online; students received a small amount of course credit for its completion. The majority of the students were White (75.5%). The data were randomly split into two groups; 1918 student records

for the Explanatory Factor Analysis (EFA) and 1926 students for the Confirmatory Factor Analysis (CFA). Only records of the students who made a good-faith effort were retained. Students were removed who incorrectly answered a marker question, who selected the same answer for all the questions, or who completed less than half of the questions. Missing data in the retained records was imputed by substituting the average for missing values.

5.2.2 Douglas Results

Douglas first removed items with an item-total or bivariate correlation which were below threshold for either exploratory or confirmatory datasets. These removals were reported in aggregate; items 4, 7, 8, 9, 18, 19, 27, 33, 38, and 41 were removed. Items with low communality coefficients were then removed leaving 24 items (the items removed by this criterion were not reported independently). Items with factor loadings less than 0.30 were removed, seven items. This left 17 items; the factor model for these 17 items was reported in the paper. The final model included 17 items composed of three factors as shown below.

1. Factor 1 - Items 3, 11, 14, 25, 28, 30, 37
2. Factor 2 - Items 5, 21, 22, 34, 40
3. Factor 3 - Items 23, 24, 29, 32, 35

Douglas then examined each factor and proposed a label for the construct it measured. Factor 1 was named Personal Application and Relation to the Real World because the items in Factor 1 relate to how students connect physics concepts to the world around them. The items in Factor 2, named Problem Solving and Learning, relate to students' attitudes toward

problem-solving and learning physics. Factor 3 was called Effort and Sense-making because the items pertain to the effort made by the students to understand physics concepts. EFA indicated that items 11, 21, 25, 34, and 35 load onto more than one factor.

Cronbach's alpha was calculated to measure the internal reliability of each factor and the overall instrument. Personal Application and Relation to the Real World had $\alpha = 0.80$, Problem Solving and Learning $\alpha = 0.73$, and Effort and Sense-Making $\alpha = 0.69$. The alpha for the overall instrument was $\alpha = 0.86$. An alpha of 0.7 is considered acceptable for low-stakes testing.

The 17-item three-factor model was then examined using CFA, however, the model fit indices were not found to be within an acceptable range. The model fit indices from the Douglas paper are provided in Table 5.1. Model fit indices included in the table are Root Mean Square Residual (RMR), Goodness of Fit (GFI), Comparative Fit Index (CFI), Root Mean Square Error of Approximation (RMSEA), and Bayesian Information Criterion (BIC). The criteria for acceptable fit require CFI or GFI to be greater than or equal to 0.95; RMR and RMSEA should be less than 0.05. The initial CFA result is reported in the row "3 factors with 17 items." Note, Kline [113] suggests reporting χ^2, RMSEA, and CFI as well as SRMR (the standardized RMR). The CFI value is fairly independent of sample size and is quite poor suggesting the model is still insufficient even after allowing errors to correlate.

Douglas then examined the modification indices of the CFA model which suggested items 11 and 35 be removed. The feature of the indices that suggested this is not reported so it could either be that cross-loading the item or adjusting the covariance with other its would improve model fit. Douglas then used modification indices to make theoretically supportable changes to the CFA model for both the 17-item and the 15-item scales (removing

11 and 35) as shown in Table 5.1. Tuning both models, yielded models with RMR, RMSEA, and GFI within the range of acceptable fit; CFI was not within range. The tuned 15-item model was reported in the paper. The tuned model allowed item 25 to load on two factors and allowed three pairs of items to covary.

Model	χ^2	df	RMR	GFI	CFI	RMSEA	BIC
3 factors and 17 items	656.84*	116	0.077	0.915	0.628	0.049	936.7
3 factors and 17 items with correlated errors	436.90*	110	0.066	0.944	0.775	0.039	762.1
3 factors and 15 items	516.70*	87	0.071	0.928	0.675	0.051	776.3
3 factors and 15 items with correlated errors	323.82*	83	0.058	0.955	0.818	0.039	603.7

Table 5.1: Confirmatory Factor analysis results for factor models ($n = 1926$) as reported in Douglas *et al.* [2]. * denotes to $p < 0.0001$.

Cronbach's was calculated again to check the internal reliability of each factor: Personal Application and Relation to the Real World $\alpha = 0.82$, Problem-Solving and Learning $\alpha = 0.73$, and Effort and Sense-Making $\alpha = 0.61$.

5.3 Replication of the Douglas Procedure

We attempted to replicate the Douglas procedure to determine if a new dataset supported the revised 3-factor model or the 8-factor model originally suggested for the instrument.

5.3.1 Replication Sample

Our replication of the Douglas procedure was performed on data collected from Fall 2015 to Fall 2019 at a large land-grant university in the eastern US. The university's general undergraduate population reported ACT scores ranging from 21 to 26 (25th to 75th percentile) [197]. The undergraduate demographic composition was 80% White, 6% inter-

national, 4% African American, 4% Hispanic, 4% students reporting two or more races, 2% Asian, and other groups each with 1% or less [197].

This study includes data from the introductory calculus-based mechanics course taken by scientists and engineers. The course was led by a single instructor with expertise in PER throughout the study. A total of 6472 students attempted the CLASS from fall 2015 to fall 2019. Only students with no missing answers and who answered the marker question correctly were retained, leaving 5809 records. This sample was split into two samples, an exploratory ($N = 2905$) and a confirmatory sample ($N = 2904$). Both the samples met the threshold number of records, 1000 [104], suggested for factor analysis.

5.3.2 Replication Results

Following the Douglas procedure, items with low item-total correlation coefficients ($r_{total} < 0.2$) or low maximum bivariate correlation ($r_{max} < 0.25$) coefficients were eliminated. Table **??** shows the descriptive statistics and correlation results for 36 items of the CLASS. Items 4, 7, 9, 31, 33, and 41 were not considered following the recommendations of the CLASS authors [1]. Item 31 is the marker questions. The other items were considered by Douglas but immediately removed because of either low item-total or bivariate correlation. The low correlation items were not consistent across the two splits of the sample. Rather than modifying the confirmatory datasets, the inconsistency was eliminated by bootstrapping. The correlations were calculated 1000 times subsampling the dataset with replacement. Items 8, 18, 19, 27, and 38 were removed because they had low maximum bi-variate correlations with other items in at least 75% of the bootstrap replications. Items 8 and 18 also had consistently low item-total correlations. No items were removed for low item-total correlation

which was not already removed for low bivariate correlation. At this stage Douglas had removed items 4, 7, 8, 9, 18, 19, 27, 31, 33, 38, and 41; our replication also removed this set of items. This left 31 items.

Table 5.2 identifies two different categories of items: "forward-coded items" for the items with which experts "strongly agreed" or "agreed" and "reverse-coded items" for the items with which experts "strongly disagreed" or "disagreed".

Item	Scoring	Mean	SD	r_{max}	r_{total}	Item	Scoring	Mean	SD	r_{max}	r_{total}
$Q01$	R	2.83	1.18	0.35	0.32	$Q22$	R	2.90	1.01	0.32	0.30
$Q02$	F	3.98	1.04	0.44	0.33	$Q23$	R	3.77	1.00	0.44	0.35
$Q03$	F	3.34	1.19	0.54	0.47	$Q24$	F	3.78	1.02	0.45	0.40
$Q05$	R	3.10	1.15	0.43	0.37	$Q25$	F	2.87	1.26	0.56	0.47
$Q06$	R	3.70	1.07	0.51	0.33	$Q26$	F	4.02	0.91	0.54	0.45
Q08	R	2.25	0.95	0.09	0.26	**Q27**	R	3.27	1.17	0.20	0.17
$Q10$	R	3.78	1.03	0.41	0.29	$Q28$	F	3.63	1.05	0.56	0.50
$Q11$	F	3.67	1.02	0.53	0.40	$Q29$	R	4.06	1.05	0.51	0.37
$Q12$	R	2.35	1.20	0.24	0.31	$Q30$	F	3.84	0.98	0.63	0.50
$Q13$	R	3.64	1.08	0.45	0.33	$Q32$	R	3.75	1.05	0.56	0.34
$Q14$	F	3.26	1.16	0.54	0.47	$Q34$	F	3.45	0.98	0.59	0.43
$Q15$	F	3.64	0.94	0.49	0.34	$Q35$	R	3.75	1.08	0.57	0.38
$Q16$	F	3.65	1.09	0.48	0.32	$Q36$	F	2.88	1.14	0.33	0.32
$Q17$	R	3.38	1.08	0.43	0.32	$Q37$	F	3.40	1.15	0.48	0.45
Q18	R	3.48	1.16	0.12	0.16	**Q38**	F	3.55	1.12	0.35	0.24
Q19	F	3.54	1.12	0.25	0.21	$Q39$	F	3.69	0.93	0.44	0.35
$Q20$	R	3.57	1.08	0.49	0.38	$Q40$	R	3.57	1.08	0.61	0.43
$Q21$	R	3.24	1.19	0.54	0.43	$Q42$	F	3.71	0.92	0.54	0.38

Table 5.2: Univariate summary statistics, maximum item bi-variate correlation and item-total correlations ($n = 2904$). Bold items 8, 18, 19, 27, and 38 were removed before the next step in developing the scales due to their maximum bi-variate correlation with other items ($r < 0.275$). The R in the column Scoring refers to the reverse-coded items and F refers the forward-coded items.

EFA was conducted first to determine the dimensionality of CLASS. Oblique rotation was used because the factors could be correlated. A scree plot was used to find the optimal number of latent factors. Figure 5.1 shows the Scree plot for the remaining 31 items in the CLASS. In the figure, the x-axis represents the number of factors, and the y-axis represents

the eigenvalues for each factor structure. The factoring algorithm finds factors as eigenvectors of the correlation matrix. All elements on the diagonal of this matrix are one; as such, the trace of the matrix is the number of items, 31. The trace of this matrix is also the sum of the eigenvalues. A factor with an eigenvalue less than one explains less variance than a single item and should not be included. The plot suggests the three factors are appropriate (two additional factors have eigenvalues slightly above one). The maximum change in curvature of the plot, known as the "knee" [108], was also observed at two to three components. A parallel-analysis curve was also examined. Parallel analysis factors a random dataset the same size as the sample dataset [217]; the optimal factors should be above the parallel analysis line. This also suggests 3 factors.

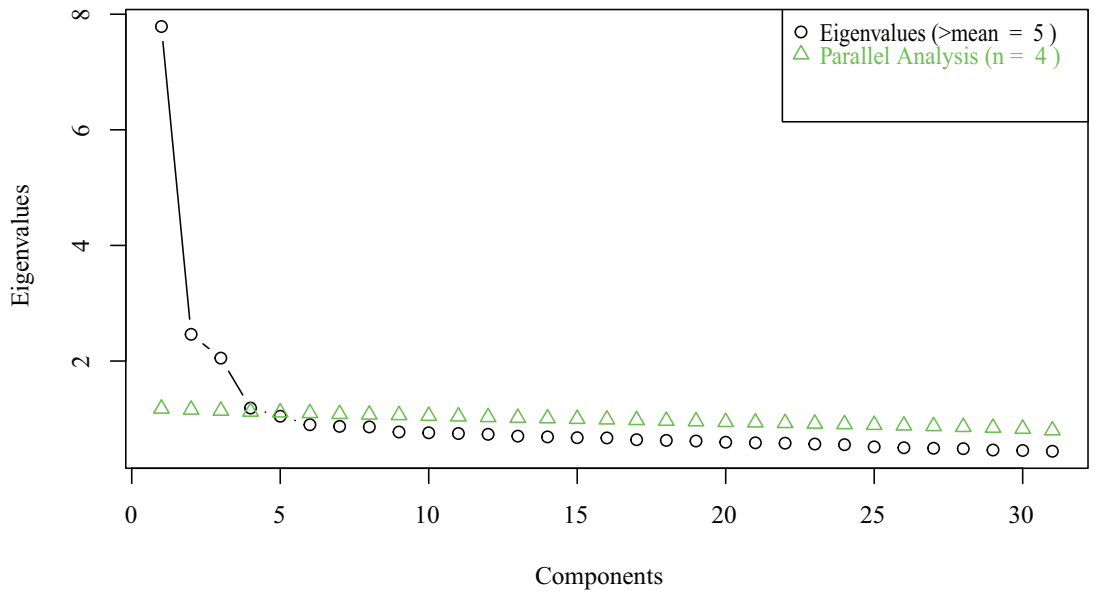


Figure 5.1: Scree plot for a 3-factor model of the items on the CLASS.

The results of the scree plot in Fig. 5.1 strongly suggest a 3-factor model for the

CLASS, which differs from the initial 8-factor model described by the CLASS authors [1], while supporting the 3-factor model suggested by both Douglas *et al.* and Kontro *et al.* [2, 215].

Following the Douglas procedure, using 3 factors, the communality coefficients of the items were examined, and items with communality less than 0.3 were iteratively removed. Again, bootstrapping was used to eliminate inconsistencies are the 0.3 threshold. Thirteen items were removed for low communality (1, 2, 10, 12, 13, 15, 16, 17, 20, 22, 23, 36, and 39) which left 17 items (3, 5, 11, 14, 23, 24, 25, 26, 29, 30, 32, 34, 35, 36, 40, and 42). Note, all items removed except 35 were of low communality in at least 80% of the bootstrap replication. Of the retained items, only item 35 had low communality in a significant number of replications (42%). The other retained items were of low communality in less than 20% of the replications with 12 of the retained items having low communality in less than 10% of the replications.

Exploratory Factors Analysis was then performed on the remaining items; we report only the items with factor loadings exceeding 0.30 in Table 5.3. The factors identified in our EFA are related to those Douglas *et al.*'s study with some differences. Many of the differences result from items with fairly weak factor loadings (< 0.5). Factor 1 was identical. Factor 2 (Problem Solving/Learning) is missing an item (item 22, "If I want to apply a method used for solving one physics problem to another problem, the problems must involve very similar situations"). The third factor is fairly different from item 26 ("In physics, it is important for me to make sense out of formulas before I can use them correctly") strongly loading on the factor; this item was not included in the Douglas factoring. If we restrict the third factor to items with loadings of 0.5 and above it contains items 24, 26, and 29; two of the three items

are also contained in the four-item Douglas Effort/Sense-Making subscale. Item 35 was also in both the Douglas model and this model for Factor 3; it had weak loadings in both models. In general, the EFA supports the three-factor solution of Douglas much more strongly than the original eight-factor model of Adams *et al.*.

Item	Factor1	Factor2	Factor3
Q3	0.77		
Q5		0.60	
Q11	0.41		
Q14	0.67		
Q21		0.50	
Q24			0.50
Q25	0.63	0.33	
Q26			0.63
Q28	0.60		
Q29			0.67
Q30	0.41		0.41
Q32			0.44
Q34		0.48	
Q35			0.38
Q37	0.53		
Q40		0.65	
Q42			0.32

Table 5.3: Exploratory Factor Analysis results replicating the EFA of Douglas *et al.* for the remaining items. Only factor loadings exceeding 0.30 are reported.

Confirmatory Factor analysis (CFA) [99] was then conducted using Structural Equation Modeling (SEM) [112] to characterize the proposed factor structure. The factor model had a GFI of 0.932, an RMR of 0.061, an RMSEA of 0.068, and a CFI of 0.887; all were outside of the range of good model fit. Most were an improvement over the 17-item Douglas values. We then tried the modifications performed to make the correlated 15-factor model by removing items 11 and 35, allowing item 25 to load on two factors, and adding three additional pairs of correlations. This produced a model with a GFI of 0.949, an RMR of 0.055, an RMSEA of 0.064, and a CFI of 0.915. Some of these were better and some worse than Douglas; taken

together they do not suggest a good-fitting model.

The next chapter explores why good fitting models were not extracted for items that superficially are meaningful and well constructed and suggests a set of scales that preserve more than 15 of the original 41 items.

Chapter 6

An Optimal CLASS Model

6.1 Introduction

The results of the preceding chapter strongly suggested the original 8-factor model for the CLASS instrument of Adams *et al.* was not a good fit for the instrument. The optimal 3-factor, 15-item model proposed by Douglas *et al.* did not have a good model fit; neither did the model result from our replication of the Douglas process. The Comparative Fit Index (CFI) of the Douglas 15-item model and our replication were quite poor. We note that while Goodness of Fit Index (GFI) was in the range of good model fit for the Douglas 15-item model, this metric has been demonstrated to have positive bias for large samples and has ceased to be broadly used [218]. Further, both models do not use the majority of the items in the instrument which may prevent those who have already collected a large CLASS data set from fully understanding the implications of the results of the instrument.

6.2 Tuning to a Good Model

To construct a good fitting model, we extended the tuning process used in the CFA to examine the modification indices and make model changes until a good model fit was obtained Root Mean Squared Error of Approximation (RMSEA<0.05), Standardized Root Mean Square Residual (SRMR<0.05), and CFI>0.95. This resulted in a significantly cross-loaded model as shown in Fig. 6.1. The model fit measures were RMSEA 0.046, CFI 0.952, and SRMR 0.033 (per Kline's suggestion we moved to SRMR). The lines in the model represent the cross-loadings.

This figure offered two clues as to why the CLASS was not producing a good factor model. The first was Item 25 ("I enjoy solving physics problems") which cross-loaded on

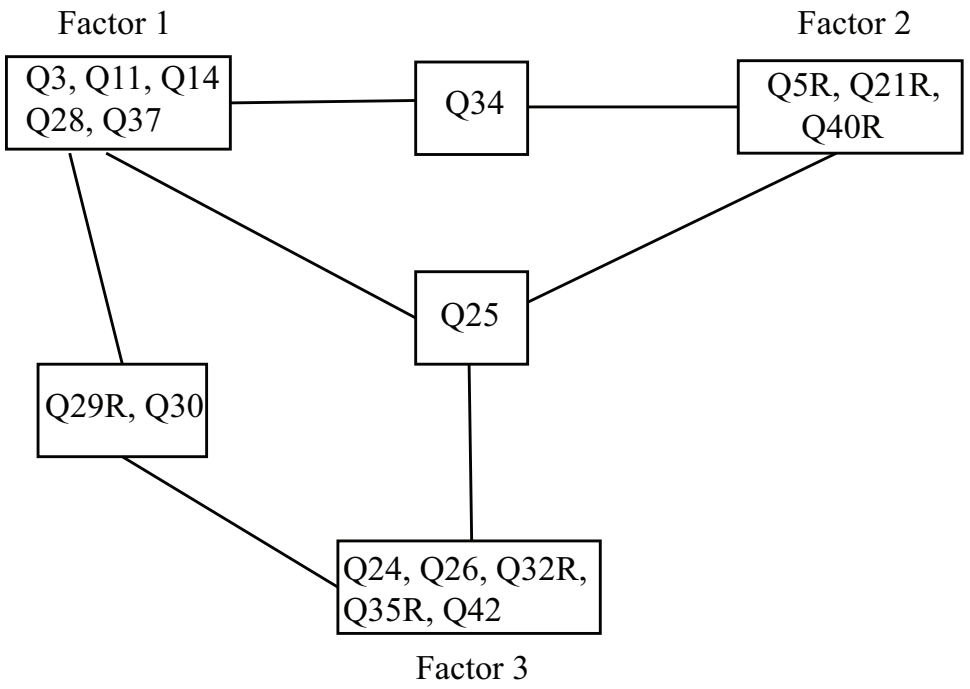


Figure 6.1: Model tuned to good model fit. Lines represent cross-loadings. Items marked with an "R" are reverse coded.

multiple factors. The other items in Factor 1 involve the connection of physics to real-world situations. While enjoying solving physics problems may result in seeing a real-world connection or maybe the result of seeing a real-world connection, it cannot be viewed as part of a unified construct of seeing a connection between physics and the real world (you certainly could have one without the other). Further, there are no other items like it in the instrument. Item 25 lacks face validity for the real-world connection construct. Item 25 also measures a valuable unique construct, enjoying solving physics problems. We identify item 25 as a "unique" item, an item where similar items were not found in the instrument, which measures a valuable construct. This item, which is theoretically related to many other constructs measured by the instrument, acts to glue multiple constructs together muddying the factor structure.

Examination of the EFA in Table 5.3 shows four items with relatively low loadings: items 11, 30, 35, and 42. Item 35 was also the retained item with the lowest communality in most bootstrap replications. Item 11 ("I am not satisfied until I understand why something works the way it does") does not mention either physics learning or physics problem-solving but is rather a general statement about student attitudes. We identify this as another "unique" construct; again other items in the instrument do not measure this construct. Similarly, item 30 makes a general statement about reasoning skills ("Reasoning skills used to understand physics can be helpful to me in my everyday life"). Superficially, this would seem like a real-world connection item, but the factor model suggests it is being interpreted more generally. We also classified Item 30 as unique.

Figure 6.1 also illustrates a second feature of the CLASS that may affect the factor structure. With the removal of Item 11, all items in Factor 1 mention the relation of physics to real life or everyday experience; as such, they have face validity for measuring a Real World Connections construct. All are also forward-coded. Factor 2 contains three items that express a lack of confidence in solving physics problems; they have face validity as measuring a reverse-coded Physics Problem-Solving Self-Efficacy construct. The third factor contains three items that mention understanding physics formulas (items 24, 26, and 32) and one item about studying and using methods beyond memorization, Item 42. The last item, Item 35 ("The subject of physics has little relation to what I experience in the real world.") is superficially a real-world connection item, but it does not factor with the other items in Factor 1. Factors 1 and 2 may explain this; Factor 1 is all forward coded while Factor 2 is all reverse coded. Item 35 fits with Factor 1 but is reverse-coded. It appears that forward-coded items are preferentially factoring with forward-coded items while reverse-coded items

preferentially factor with reverse-coded items. With this observation, we added Item 35 to Factor 1 and will determine if it fits in this factor using a reliability analysis (Cronbach's alpha) and CFA. Using the same reasoning, item 34 ("I can usually figure out a way to solve a physics problem") is a straightforward forward-coded statement of problem-solving self-efficacy (and not about real-world connections); we tentatively added this item to Factor 2. This reasoning produced two coherent subscales: items 3, 14, 28, 35, and 37 (Real World Connections (RWC)) and items 5, 21, 34, and 40 (Physics Problem Solving Self-Efficacy (PPSSE)). We have identified three unique items: 11, 25, and 30.

The observation that reversed and forward-coded items were factoring differently suggested examining each item group separately. For the remaining forward-coded items (items 2, 15, 16, 24, 26, 36, 39, and 42), the scree plot suggested a one-factor solution. All items had strong loadings for the single factor except item 36 ("There are times I solve physics problems more than one way to help my understanding"). While certainly good practice, this behavior is likely something not done by most students; this item was discarded (eliminated from the set of items under consideration). The item superficially belongs with some other items in the instrument, but the factor analysis suggested it was not behaving as one would expect. Examining the remaining items, all items except Item 16 refer to good physics problem-solving or learning strategies. Item 16 ("Nearly everyone is capable of understanding physics if they work at it.") is different than the other items. It appears to be a simple statement of having a growth mindset. This item was classified as unique. The remaining items (2, 15, 24, 26, 39, and 42) were identified as an "Expert-like Physics Problem Solving" (EPPS) subscale.

For the remaining reverse-coded items (items 1, 6, 10, 12, 13, 17, 20, 22, 23, 29, and

32), the scree plots suggested a 2-factor solution. Examining the second factor, Item 22 ("If I want to apply a method used for solving one physics problem to another physics problem, the problems must involve very similar situations") had a low loading on both factors. This statement seems fairly ambiguous and it is unclear how it would be interpreted by students. This item was discarded. Item 12 ("I cannot learn physics if the teacher does not explain it well in class") was the strongest loading on the second factor. It represents an item that would likely be answered positively by most students and negatively by experts; the idea one can learn physics independent of instruction. Again, there are no other items like it in the instrument. This item was classified as unique. The factor analysis was rerun eliminating items 12 and 22. The scree plot suggested a one-factor solution. Item 1 ("A significant problem in learning physics is being able to memorize all the information I need to know") did not load on the factor and was discarded. While the previous scales extracted dealt with physics problem solving, these reverse coded items dealt more generally with the learning and understanding of physics. We named this scale Expert-like Physics Learning (EPL) (items 6, 10, 13, 17, 20, 23, 29, and 32).

Because some of the items that were not factoring properly contained valuable constructs, we returned to the items removed at the early stage of the replication process in the previous chapter. The removed items were examined, some represented valuable constructs not related to other items; these items were classified as unique. These items could be developed into interesting subscales in the future. Some items seemed related to other items and did not seem to represent a unique additional construct; these items were discarded. These items should have been correlated with other items in the instrument; they were removed as likely problematic.

- Item 8 (2.26 ± 1.0) (Discarded) - "When I solve a physics problem, I locate an equation that uses the variables given in the problem and plug in the values." It seems likely many introductory physics students would naturally answer positively.

- Item 18 (3.48 ± 1.2) (Discarded) -"There could be two different correct values to a physics problem if I use two different approaches." This would appear to be a naïve problem-solving item but does not correlate with other such items.

- Item 19 (3.53 ± 1.1) (Unique) -"To understand physics I discuss it with friends and other students" - It is not clear if there is a correct answer, but it seems useful to know this about a student.

- Item 27 (3.29 ± 1.2) (Unique) - "It is important for the government to approve new scientific ideas before they can be widely accepted." This is unlike any other item in the instrument.

- Item 38 (3.54 ± 1.1) (Discarded) -"It is possible to explain physics ideas without mathematical formulas" This one has many possible meanings and may be answered inconsistently even by expert physicists. It is also different than other items in the instrument.

6.3 Final Subscales

Building on the preliminary EFA results and the scoring of the items (reverse or forward), four distinct subscales were identified. Each subscale underwent individual testing using CFA to evaluate model fit and Cronbach's alpha to evaluate reliability. Multiple

statistical parameters were examined to assess the absolute fit and comparative fit. The Comparative Fit Index (CFI) and Tucker-Lewis Index (TLI) were calculated as comparative model fit parameters. Standardized Root Mean Square Residual (SRMR) and Root Mean Square Error of Approximation (RMSEA) were used to test absolute model fit [113, 2].

Factor	Items	df	p	CFI	TLI	RMSEA	SRMR	α
RWC	3, 14, 28, 35, 37	5	0.002	0.995	0.990	0.031	0.014	0.75
PPSSE	5, 21, 34, 40	4	0.000	0.989	0.967	0.062	0.018	0.69
EPPS	2, 15, 24, 26, 39, 42	6	0.000	0.990	0.983	0.034	0.018	0.74
EPL	6, 10, 13, 17, 20, 23, 29, 32	8	0.000	0.980	0.972	0.037	0.022	0.76

Table 6.1: Confirmatory Factor analysis results for final scales ($n = 2904$). Factors are named with related to the work of Adam *et al.* and Douglas *et al.* Real World Connection (RWC), Physics Problem Solving Self-Efficacy (PPSSE), Expert-like Physics Learning (EPL), and Expert-like Physics Problem Solving (EPPS). Model parameters included in the table are the comparative Fit Index (CFI), Tucker-Lewis Index (TLI), Root Mean Square Error of Approximation (RMSEA), Standardized Root Mean Square Residual (SRMR), and Cronbach's α.

In the Adam *et al.* study, three items (3, 14, and 28) were categorized under the factor "Personal Interest", while three items (28, 35, and 37) were grouped under the factor "Real World Connections". These six items were combined to form the first subscale RWC. In the Douglas *et al.* study, the same four items (3, 14, 28, and 37) from the RWC subscale were part of a factor called "Personal Application and Real World Connections". In our RWC scale, only item 35 was reverse coded, while all other items were forward coded. The use of reverse-coded items allows for a balanced representation of responses and mitigates potential response biases; however, when the constructs in the instrument are closely related, the differential response rates to positive and negative items may become a problem. The CFA results for RWC, presented in Table 6.1, demonstrated good fit statistics. The Cronbach's alpha coefficient of 0.75 is above the threshold of 0.7 for a reliable scale for low-stakes testing.

The second subscale, "Physics Problem Solving Self Efficacy (PPSSE)", contained

items 5, 21, 34, and 40. These items are loaded onto many different factors in the Adams *et al.* model. The four items also belonged to the factor 'Problem Solving/Learning' in the Douglas *et al.* study. Only Item 34 was forward-coded. CFA results (Table 6.1) generally demonstrate good model fit statistics with alpha just below the 0.7 threshold. This is likely because of the small number of items on the scale. The RMSEA was above the threshold of good model fit (0.05) but below the threshold for poor model fit (0.1).

The third subscale, "Expert-like Physics Problem Solving (EPPS)" contained items 2, 15, 24, 26, 39, and 42. These items loaded on many different factors in Adams *et al.* with Item 2 not loading onto any factor. Item 24 belonged to the factor "Effort/Sense Making" subscale in the Douglas *et al.* study, but there was generally little relation between this subscale and the Douglas factors. All 8 items were forward-coded. CFA results (Table 6.1) indicate good model fit and reliability.

The fourth subscale, "Expert-like Physics Learning (EPL)" contained items 6, 10, 13, 17, 20, 23, 29, and 32. These items also belonged to many Adams *et al.* factors with items 10, 17, 20, and 29 not loading onto any factor. Three items, 23, 29, and 32 also belonged to the factor 'Effort/Sense Making' in the Douglas *et al.* study. All 8 items were reverse-coded. This subscale also had good model fit and reliability (Table 6.1).

6.4 Discussion

The above analysis suggests a four-subscale model for the CLASS. In addition to the 12 items 3, 5, 14, 21, 23, 24, 28, 29, 32, 34, 37, and 40 included in the Douglas *et al.* subscales, 11 extra items 2, 6, 10, 13, 15, 17, 20, 26, 35, 39, and 42 were included in the new subscale

structure. Item 22, "If I want to apply a method used for solving one physics problem to another problem, the problems must involve very similar situations" in the Douglas factor Problem Solving/Learning excluded from the subscale structure due to the low communality.

The Real World Connections subscale contained items related to how students attempt to connect or combine the physics they learn in the classroom with their everyday lives. The RWC items follow:

- Item 3 - "I think about the physics I experience in everyday life."
- Item 14 - "I study physics to learn knowledge that will be useful in my life outside of school."
- Item 28 - "Learning physics changes my ideas about how the world works."
- Item 35 - "The subject of physics has little relation to what I experience in the real world."
- Item 37 - "To understand physics, I sometimes think about my personal experiences and relate them to the topic being analyzed."

Out of the five items, four items 3, 14, 28, and 37 were also included in the Douglas *et al.* factor Personal Application and Relation to Real World. Item 35, "The subject of physics has little relation to what I experience in the real world.", was not included in the Douglas *et al.* study; however, the statement emphasizes the relation of the real world and physics content. The reverse coding of the item may have caused it to factor separately from the other items. The subscale had a good model fit showing the inclusion of item 35 was warranted.

The second subscale, Physics Problem Solving Self Efficacy, collected items related to students' beliefs about their ability to solve physics problems. The PPSSE items follow:

- Item 5 - "After I study a topic in physics and feel that I understand it, I have difficulty solving problems on the same topic."
- Item 21 - "If I don't remember a particular equation needed to solve a problem on an exam, there's nothing much I can do (legally!) to come up with it."
- Item 34 - "I can usually figure out a way to solve physics problems."
- Item 40 - "If I get stuck on a physics problem, there is no chance I'll figure it out on my own."

The only difference between this subscale from the factor 'Problem Solving/Learning' in the Douglas *et al.* study was Item 22, "If I want to apply a method used for solving one physics problem to another problem, the problems must involve very similar situations" was not included in the our subscales. Item 34 is forward-coded while items 5, 21, and 40 are reverse-coded. The four items had a generally good fit model with RMSEA above the range of good fit and alpha slightly below the threshold for good reliability.

Six CLASS items related to how students approach a physics problem or their thought process while solving a physics problem were collected into a third subscale, "Expert-like Physics Problem Solving". This scale differs from the previous subscale, PPSSE, because the items discuss general features of physics problem solving, not the students' beliefs of their ability to solve problems. The six items included are:

- Item 2 - "When I am solving a physics problem, I try to decide what would be a reasonable value for the answer."
- Item 15 - "If I get stuck on a physics problem on my first try, I usually try to figure out a different way that works."
- Item 24 - "In physics, it is important for me to make sense out of formulas before I can use them correctly."
- Item 26 - "In physics, mathematical formulas express meaningful relationships among measurable quantities."
- Item 39 - "When I solve a physics problem, I explicitly think about which physics ideas apply to the problem."
- Item 42 - "When studying physics, I relate the important information to what I already know rather than just memorizing it the way it is presented."

All items are forward-coded. Out of these six items only item 24 was included in the Douglas *et al.* Effort/Sense Making factor. Model parameters calculated in CFA supported good model fit.

The fourth subscale, "Expert-like Physics Learning" contains 8 items. Three items belonged to the Effort/Sense Making factor in the Douglas *et al.* study (items 23, 29, and 32). Five items (items 6, 10, 13, 17, and 20) were not included in the Douglas factors. The EPL items follow:

- Item 6 - "Knowledge in physics consists of many disconnected topics."

- Item 10 - "There is usually only one correct approach to solving a physics problem."
- Item 13 - "I do not expect physics equations to help my understanding of the ideas; they are just for doing calculations."
- Item 17 - "Understanding physics means being able to recall something you've read or been shown."
- Item 20 - "I do not spend more than five minutes stuck on a physics problem before giving up or seeking help from someone else."
- Item 23 - "In doing a physics problem, if my calculation gives a result very different from what I'd expect, I'd trust the calculation rather than going back through the problem."
- Item 29 - "To learn physics, I only need to memorize solutions to sample problems."
- Item 32 - "Spending a lot of time understanding where formulas come from is a waste of time."

Lower scores or disagreement with the items in this subscale represented the expert-like physics learning attitudes; agreement with these items shows naive physics student beliefs. In the Douglas *et al.* study, the latter three items were grouped with item 24 ("In physics, it is important for me to make sense out of formulas before I can use them correctly.") in the Effort/Sense Making factor. Grouping these eight items to form a subscale appears to be justified as the model parameters calculated in the CFA consistently indicated a good fit.

EFA and CFA statistics supported a subscale structure with four latent dimensions: Real World Connection, Physics Problem Solving Self-Efficacy, Expert-like Physics Learning,

and Expert-like Physics Problem Solving using 23 items of the CLASS. Table 6.2 presents the final subscale structure.

	Items
RWC	3, 14, 28, 35, 37
PPSSE	5, 21, 34, 40
EPPS	2, 15, 24, 26, 39, 42
EPL	6, 10, 13, 17, 20, 23, 29, 32
Unique Items	11, 12, 16, 19, 25, 27, 30
Discarded Items	1, 8, 18, 22, 36, 38
Not scored Items	4, 7, 9, 31, 33, 41

Table 6.2: Final factor structure developed. Factors are named related to the work of Adam *et al.* and Douglas *et al.* , Real World Connection (RWC), Physics Problem Solving Self-Efficacy (PPSSE), Expert-like Physics Learning (EPL), and Expert-like Physics Problem Solving (EPPS).

In addition to the four subscales, a fifth category, unique items, was created: items 11, 12, 16, 19, 25, 27, and 30. These items either stood out in the statistical process as different than other items in the instrument or were identified as the final scales were explored for face validity. Item 11, "I am not satisfied until I understand why something works the way it does" and item 30, "Reasoning skills used to understand physics can be helpful to me in my everyday life " had weak factor loadings on both the RWC and EPL subscales; adding the two items into either subscale did not improve the model fit. Item 11 makes a general statement not specific to physics. Item 30 specifically involves real-world connections but statistics suggest it is being interpreted differently than other items in the RWC scale. Item 12, "I cannot learn physics if the teacher does not explain things well in class" loaded onto EPL; however, the model parameters were not improved when added into the subscale as the item was not strongly correlated to the rest of the items in the subscale. This seems an item that most physics students would agree with. Item 25, "I enjoy solving physics problems" significantly loaded into the three subscales RWC, PPSSE, and EPPS;

model fit was not improved by adding the item into any of the subscales. The statement is completely different than all other items in the instrument. Item 16, "Nearly everyone is capable of understanding physics if they work at it", is significantly loaded in the EPPS subscale; however, it is also very different than other items in the scale. This item represents the attitude of a grown mindset. Item 19, "To understand physics I discuss it with friends and other students" weakly loaded onto the EPPS subscale. It is not clear whether there is an expert-like response to this item. Item 27, "It is important for the government to approve new scientific ideas before they can be widely accepted" weakly loaded onto the EPL subscale. It also is completely different than all other items in the instrument.

CFA was repeated for the final four-scale model (excluding the unique items) combining all items in all subscales which did not produce a good fitting model (it was not expected to) with RMSEA of 0.052, SRMR of 0.050, and CFI of 0.889. This model is shown in Fig. 6.2. The RMSEA and SRMR were at the threshold of good model fit; the CFI was below the threshold but better than that of the Douglas model. No fit statistics were reported for the Adams model; in fact, none are calculable because of the massively cross-loaded nature of the model.

As before, the model was then tuned using modification indices producing the model in Fig.6.3. This required cross-loading items 32, 34, and 35 and allowing four pairs of correlations beyond that explained by the factor. The tuned model had RMSEA of 0.036, SRMR of 0.031, and CFA of 0.95. If items 32, 34, and 35 are removed a good fitting model can be constructed with no cross-loadings by allowing three pairs of items to covary yielding a model with RMSEA of 0.037, SRMR of 0.031, and CFA of 0.95. Removing items 32, 34, and 35 reduces the internal reliability of the RWC, PPSSE, and EPL subscales; these items

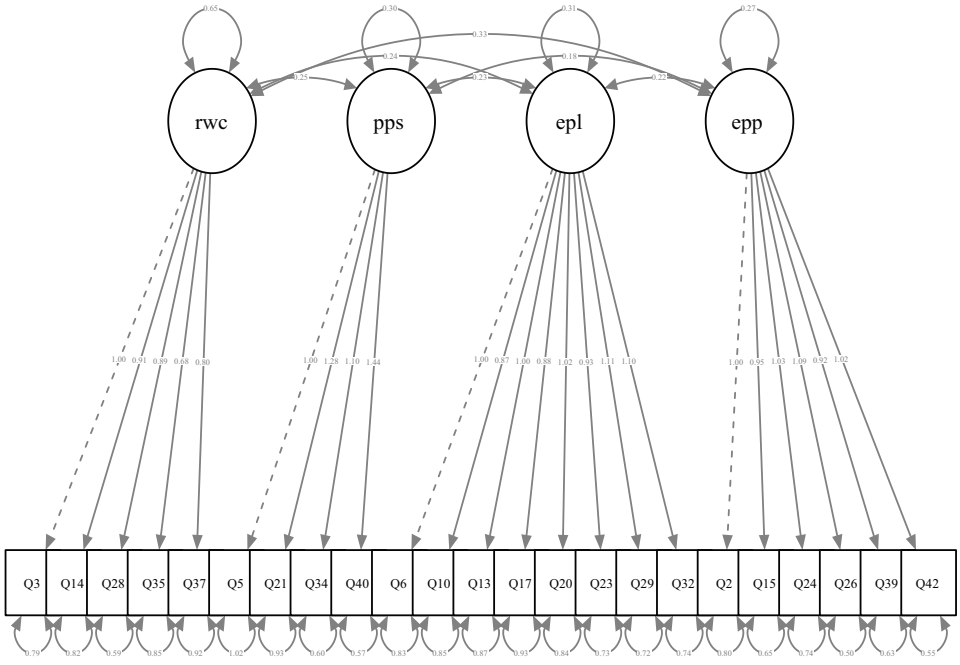


Figure 6.2: Unadjusted SEM for final four-subscale model.

also superficially belong in these scales. As such, their inclusion seems to be warranted even with the necessity of cross-loading the model.

The EFA and CFA conducted in this study, as well as general theoretical considerations, suggest that the optimal subscale structure is not a simple three or four-subscale model, but instead, a more complex configuration. The new subscales as well as the unique items allow for the exploration of a rich set of models. A basic property of an SEM model is that any covariance can be turned into a regression without changing the model fit. Figure 6.4 shows a model that proposes that EPL, EPPS, and PPSSE cause RWC while EPL and EPPS cause PPSSE. Naturally, this causal hypothesis would have to be tested by additional research.

In conclusion, the EFA and CFA results underscore the complexity and richness of the instrument, indicating the presence of intricate interactions among the items and subscales. The development of non-orthogonal subscales suggests the existence of latent variables that

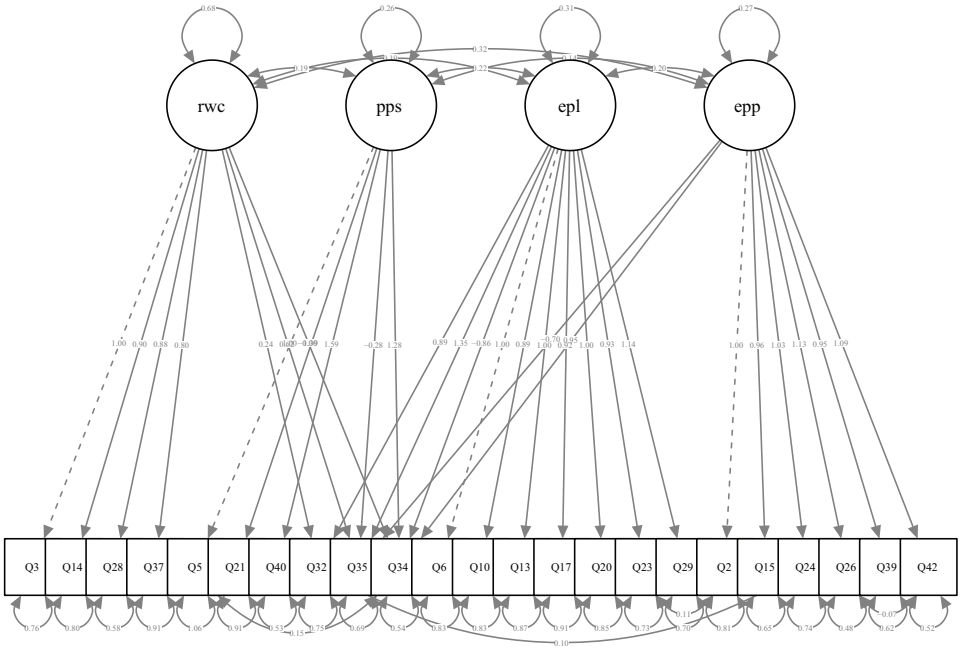


Figure 6.3: Adjusted SEM for a final four-subscale model with cross-loading and additional covariance.

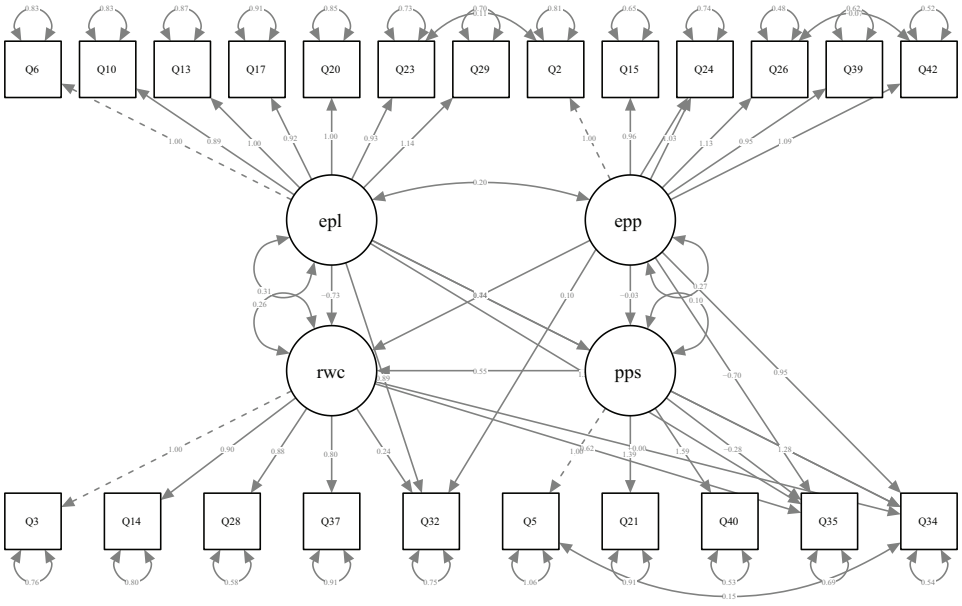


Figure 6.4: Adjusted SEM for the final four-subscale model using theoretical regression model instead of covariances.

warrant deeper exploration. This study lays the foundation for future research to gain a more comprehensive understanding of the relationships between the variables in the research domain.

Chapter 7

Exploring the Relation of Personality and Self-Efficacy on College Achievement

*

*Parts of this chapter were published in "Henderson, R., Hewagallage, D., Follmer, J., Michaluk, L., Deshler, J., Fuller, E., & Stewart, J. (2022). *Mediating role of personality in the relation of gender to self-efficacy in physics and mathematics.* Physical Review Physics Education Research, **18(1)**, 010143."

7.1 Introduction

According to the National Center for Education Statistics, both men and women enroll in high school STEM courses at the same rate, including physics [219]. They also enroll and complete undergraduate education at approximately the same rate with women at a slightly higher rate than men. The enrollment of women in physical sciences, engineering, and mathematics is substantially lower compared to men. PER has long explored differences in achievement between demographic groups. Early research explored differences in men's and women's conceptual understanding [77, 220]. Recent studies have examined other demographic groups [126, 221–223]. The relationship between non-cognitive factors and academic outcomes is an active research area in general educational research. Substantial relations between a number of non-cognitive factors and academic outcomes have been identified [167]. This study attempts to explore the relations between two of the most studied non-cognitive factors: self-efficacy and the five-factor model of personality. Multiple research studies investigate the differences in self-efficacy of men and women; however, few studies have explored differences in personality by gender and the relationships among gender, self-efficacy, and personality [224].

Self-Efficacy in Physics

Recent research has explored the relationship between self-efficacy and physics instruction. Nissen analyzed students' self-efficacy in high school classrooms. Men reported higher self-efficacy toward physics than women and, after physics instruction, the self-efficacy of women decreased more than men [224]. Marshman *et al.* [225] also found that women re-

ported lower levels of self-efficacy in general, as well as lower values associated with physics, after completing an introductory physics course.

Significant gender differences in the self-efficacy of non-STEM majors have also been reported in physics classes. Cavallo, Rozman, and Potter showed that men reported higher self-efficacy than women in an introductory, algebra-based physics course for students majoring in the biological sciences [226]. Lindstrøm and Sharma investigated differences in self-efficacy in a course designed for students without any prior physics instruction; the gender differences in self-efficacy grew from pre-instruction to post-instruction [227].

Other studies have explored gender differences in physics classrooms required for students majoring in STEM. Women, on average, reported lower self-efficacy toward physics-related activities [228, 177, 131, 229]. In general, from pretest to post-test, women's self-efficacy toward physics decreased more than men's which, in turn, increased the gender gap in self-efficacy over the course of the semester. Furthermore, Sawtelle, Brewe, and Kramer showed that traditional instruction negatively impacted all students' self-efficacy but students' engagement in a Modeling Instruction course positively impacted specific sources of women's self-efficacy [178]. Specifically, the vicarious learning source of self-efficacy was positively related to the success of women in these courses; however, the mastery experience and social persuasion sources did not explain any variation in their models. In addition to Modeling Instruction, Miller *et al.* showed that in a class using Peer Instruction, the lower self-efficacy reported by women prior to instruction fully mediated the effect of women switching their answers to conceptual physics problems from right to wrong [230].

More recently, researchers have explored gender differences in self-efficacy in relation to how students perceive their own class achievement [231] and what harm these differences have

on female students [232]. In general, men and women interpret grades and other measures of STEM performance differently, which may lead to a confidence gap between men and women [156] and, thus, longer-term impacts such as persistence toward a degree or career.

Self-Efficacy in mathematics

In addition to studies done within the context of the physics classroom, researchers have investigated students' self-efficacy in mathematics courses. In a meta-analysis, Huang demonstrated that men reported higher self-efficacy than women toward college mathematics (Hedge's $g = 0.18$) - a similar result was reported for both physics and computer science [233]. The effect size criteria for Hedge's g are similar to Cohen's d; 0.2 is a small effect, 0.5 is a medium effect, and 0.8 is a large effect. While the results in physics show men consistently reporting higher levels of self-efficacy toward the discipline than women, in mathematics, some studies do not find significant differences in self-efficacy between men and women [234, 235]. Further, an in-depth examination of students' self-efficacy in mathematics contexts suggests that students' efficacy beliefs fluctuate over time and across engagement with mathematics tasks [236].

Self-Efficacy in other STEM disciplines

Self-efficacy has also been studied in other STEM disciplines, including engineering, biology, and chemistry. In engineering, many studies concluded that men, on average, report higher self-efficacy than women [237–239, 176, 240, 241, 201, 242–244]. Fewer studies have investigated self-efficacy in chemistry and biology. One study of high school students showed no difference in self-efficacy toward biology [245]; however, Ainscough *et al.* demonstrated

that men reported a higher level of self-efficacy toward biology at the beginning and end of a first-year biology course [246]. In college chemistry, some studies have shown that women report higher self-efficacy toward chemistry tasks than men [247, 248].

Self-Efficacy and personality

A number of studies have explored the relationship between personality and self-efficacy [249, 250], with some studies exploring differences between men and women. In a prior study, the five personality facets were measured to have different regression coefficients, β, predicting self-efficacy: neuroticism $\beta = -0.25$, extraversion $\beta = 0.27$, openness $\beta = 0.13$, agreeableness $\beta = -0.06$, and conscientiousness $\beta = 0.18$ [251]. Self-efficacy has been shown to mediate the relationship between study engagement and openness to experience [252].

A significant positive interaction between gender and emotional stability (reverse-coded neuroticism) was reported in these two variables' effect on self-efficacy [253]. Further, in a meta-analysis examining the correlates with complex task performance in the workplace, both conscientiousness and cognitive ability were positively correlated with complex task performance; conscientiousness was also positively correlated with self-efficacy ($r = 0.27$). Self-efficacy mediated the relation between conscientiousness, cognitive ability, and simple task performance but was not a mediator for complex task performance [254].

Studies have also examined links among students' self-regulation and self-efficacy skills, personality, motivation, and achievement. In an examination of profiles of college students, Dörrenbächer and Perels found that students' low or high in self-regulated learning varied in dimensions of motivation and personality [255]. Specifically, achievement was found to be significantly higher among students high in both self-regulated learning and motiva-

tion. Importantly, students higher in self-regulated learning skills, such as self-efficacy, also demonstrated lower neuroticism and higher conscientiousness, agreeableness, and openness to experience, suggesting key links among students' regulation, efficacy, and dimensions of personality.

7.1.1 Research questions

The relationships among gender, personality, self-efficacy, and physics course grade are explored within the framework of mediation and moderation. One variable mediates the relation between two other variables if part of the effect of one variable on the other is explained by the mediating variable. A variable moderates the relation between two variables if the relation is different depending on the value of the moderator. Mediation and moderation are explained in detail in Sec. 7.2.3. This study seeks to answer five research questions.

RQ1: Does self-efficacy or personality differ for men and women in core university introductory mathematics and physics classes?

RQ2: Does personality mediate the relationship of gender to self-efficacy? If so, how does it mediate the relationship?

RQ3: Does personality mediate the relationship of gender to achievement? If so, how does it mediate the relationship?

RQ4: Does self-efficacy mediate the relationship of personality and gender to achievement? If so, how does it mediate the relationship?

RQ5: Does gender moderate the relationships of personality, self-efficacy, and achievement?

Considerable efforts have been directed toward understanding and mitigating performance differences between students who are traditionally underrepresented in Science, Technology, Engineering, and Mathematics (STEM) and those who are not. The use of cognitive factors such as high school GPA, ACT, or SAT scores, conceptual pretest or post-test scores, and test averages are prominent in PER. Less work has investigated non-cognitive factors in PER - factors that do not directly measure academic achievement at some level. Richardson, Abraham, and Bond [167] provided an extensive overview of the relations between non-cognitive factors and academic achievement in the broader educational research literature.

The majority of this work has examined characteristics of instruction as one factor that may co-explain disparities in representation and performance in physics [225]. Much less work has focused on the roles of students' characteristics in explaining female students' pursuit of a STEM major or a STEM degree. Non-cognitive factors could be particularly important for understanding the under-representation of some groups in STEM [256] as well as the retention of more students within various STEM majors. Self-efficacy has long been an important variable in models predicting college persistence [257].

7.1.2 Science and mathematics anxiety

Mathematics anxiety [185, 184] and science anxiety [258, 259, 186, 188] explain a substantial amount of differences in quantitative examination performance. A meta-analysis conducted by Ma reported a negative correlation between higher levels of mathematics anxi-

ety and performance ($r = -0.27$) independent of gender [185]. Differences between men and women in mathematics anxiety ($d = 0.28$) and self-efficacy ($d = 0.33$) have the same effect size, characterized by Cohen's d, which were substantially larger than academic performance differences ($d = 0.11$). Science anxiety is lower in STEM majors than non-science majors [187]; however, within STEM, women report higher levels of science anxiety than men.

Studies to explore the sources of anxiety within the physics classroom reported students with higher communication apprehension produced lower normalized gains on the Force Concept Inventory [21, 260] and students in classes where the physics instructor allowed more autonomy had less anxiety about the course and had higher achievement in the course [261].

7.1.3 Theoretical Framework

In this research, we adopt Bandura's Social Cognitive Theory (SCT) as the primary theoretical framework for this study [169]. SCT proposes a recursive relationship between task achievement, goal setting, and self-efficacy, producing a construct that evolves in time due to external feedback and influences how an individual addresses future goals. For academic self-efficacy, one of the primary sources of performance feedback is academic achievement in classes. For STEM students, performance on course examinations often forms a substantial part of overall course grades. Substantial literature suggests experiencing stress or anxiety during an examination degrades performance; thus, the personality facet neuroticism may be related to examination performance and affect self-efficacy by modifying examination performance. Science and mathematics anxiety represent anxiety specifically experienced during mathematics and science experiences, often examinations, and are different from the general

tendency to feel anxiety measured by the neuroticism facet. However, it seems reasonable to hypothesize that students who are more likely to feel anxiety, in general, are also more likely to feel anxiety toward mathematics and science experiences.

Conversely, additional anxiety might make a student prepare for the examination more thoroughly, potentially increasing performance and, later, self-efficacy. Course grades in STEM classes also generally require the completion of assignments, such as homework, in addition to examinations. The conscientiousness facet may influence whether a student consistently completes assigned tasks, thus affecting homework and other assignment grades. Homework is generally designed to affect test performance, further suggesting conscientiousness may influence overall class grades.

Self-efficacy is a central component of Bandura's Social Cognitive Theory [169] as well as Lent, Brown, and Hackett's Social Cognitive Career Theory [262]. These particular theories rely on recursive relationships where past successes and failures inform current self-efficacy and therefore performance decisions which then influence future successes or failures. As an example of how self-efficacy is related to this recursive structure, Diseth demonstrated that self-efficacy mediates the relationship between past academic achievement and present academic achievement [263].

Developing a model for the relation of self-efficacy to academic achievement is challenging because a college STEM student's self-efficacy towards STEM academic situations has been under development for a decade before they enter an introductory physics or mathematics class. According to Bandura's model, a student's current self-efficacy should be informed by their history of both prior academic achievement and prior levels of self-efficacy as self-efficacy is adjusted according to performance feedback. As such, current self-efficacy

is partially a measure of past academic achievement which naturally affects future grades. Self-efficacy at a certain time is then properly modeled as a student's current belief in their capability to perform some action as well as their interpretation of the past experiences that inform that belief.

The connections between gender, personality, self-efficacy, and achievement are complex as detailed above. Of the myriad relations discussed earlier, the following most directly impact this work. Many studies have identified differences in self-efficacy by gender in STEM fields. Self-efficacy is a strong correlate to academic achievement. Gender differences in personality have also been identified in a very large non-STEM study. Personality facets, particularly conscientiousness, also correlate with academic achievement. Multiple studies have shown a negative correlation between STEM achievement and the tendency to feel anxiety in STEM testing situations.

7.2 Methods

7.2.1 Sample

This study was performed from Fall 2015 to Fall 2019 at a large land-grant university in the eastern United States. The university's general undergraduate population reported ACT scores ranging from 21 to 26 (25th to 75th percentile) [197]. The undergraduate demographic composition was 80% White, 6% international, 4% African American, 4% Hispanic, 4% students reporting two or more races, 2% Asian, and other groups each with 1% or less [197].

This study includes data from two introductory physics classes (Physics 1 and Physics

2) and three introductory calculus classes (Calculus 1, Calculus 1A, and Calculus 1B). Calculus 1 is the traditional one-semester Calculus 1 class. Science and engineering students who are "math ready" enroll in Calculus 1 in their first semester. Calculus 1A and Calculus 1B is a two-semester sequence that covers the material of Calculus 1 along with some pre-calculus and is designed for students who are not academically ready to take Calculus 1. Physics 1 is the introductory calculus-based mechanics course taken by scientists and engineers and has Calculus 1 or Calculus 1A as its prerequisite. Physics 2 is the introductory calculus-based electricity and magnetism course and has Physics 1 as its prerequisite. Data were also collected in Workshop Mathematics, a remedial mathematics class to help students prepare to take college algebra. Over 90% of the students in the calculus and physics classes were pursuing STEM majors, while only 70% of the students in Workshop Mathematics were pursuing STEM majors. Workshop students represent a population more representative of the university in general.

The mathematics classes were taught by many instructors over the course of the study. Many of these instructors used a variety of active learning strategies to support their students. Each physics course was led by a single instructor with expertise in PER over the course of the study. The lectures used clicker questions and research-based pedagogy to engage students. Multiple other instructor teams taught the courses and adopted this pedagogy; the class used multiple lecture sections per semester. Both physics classes had a 3-hour per week required lab which used small group problem solving, whiteboarding, and hands-on inquiry-based activities.

A total of 6286 students completed the physics classes from the Fall 2015 to Fall 2019 and a total of 8937 students completed the mathematics classes from fall 2016 to spring

2019. Of these, only domestic students with ACT or SAT scores who completed both the personality and self-efficacy surveys were retained. For physics, 3334 students met these criteria (1783 Physics 1 and 1551 Physics 2). For mathematics, 3977 students met these criteria (1074 Workshop Mathematics, 765 Calculus 1A, 563 Calculus 1B and 1575 Calculus 1) completed both surveys. These students form the sample for this study.

7.2.2 Instruments

Big Five Inventory - BFI

The Big Five Inventory (BFI) [264, 180–182] was used to measure students' personality. The BFI measures the five-factor personality model based on the following facets: agreeableness, conscientiousness, extraversion, neuroticism, and openness. The BFI has been extensively used in a broad set of studies [198]. This work focuses on the conscientiousness and neuroticism facets.

Neuroticism is related to the tendency to feel stress, anxiety, or other strong emotions. The BFI measures neuroticism using eight items measured on a five-point Likert scale. The student is asked to rate how true the statement is for them; some items are reversed coded. The items are:

- Is depressed, blue
- Is relaxed, handles stress well (reversed)
- Can be tense
- Worries a lot

- Is emotionally stable, not easily upset (reversed)
- Can be moody
- Remains calm in tense situations (reversed)
- Gets nervous easily

Conscientiousness is related to the tendency to follow instructions, to work hard and carefully, and to meet outside expectations. The items in the conscientiousness scale in the BFI are:

- Does a thorough job
- Can be somewhat careless (reversed)
- Is a reliable worker
- Tends to be disorganized (reversed)
- Tends to be lazy (reversed)
- Perseveres until the task is finished
- Does things efficiently
- Makes plans and follows through with them
- Is easily distracted (reversed)

Motivated Strategies for Learning Questionaire - MSLQ

Self-efficacy was measured with the Self-Efficacy for Learning and Performance subscale from the Motivated Strategies for Learning Questionnaire (MSLQ) [199]. This eight-item scale was reduced to six items and specialized to the class environment by specifying either physics or mathematics classes [265]. The scale was reduced to remove one item asking about reading comprehension and one item that was very similar to a second item as part of a larger project to measure self-efficacy in multiple STEM domains. The resulting physics self-efficacy subscale is:

- I believe that I will receive an excellent grade in this physics class.
- I'm confident I can understand the basic concepts taught in this physics class.
- I'm confident I can understand the most complex material presented by the instructor in this physics class.
- I'm confident I can do an excellent job on the assignments and tests in this physics class.
- I'm certain I can master the skills being taught in this physics class.
- Considering the difficulty of this course, the teacher, and my skills, I think I will do well in this physics class.

The modified survey items were extensively revalidated. The method of word substitution to modify the MSLQ for specific domains has been used in prior studies [201].

Both of the surveys were administered once per semester and the students received a small amount of course credit for completing each survey. Informed consent was collected from all participants and all procedures were approved by the Institutional Review Board.

Academic achievement was characterized by physics and mathematics course grades measured on a numeric scale with F=0 and A=4. Academic preparation was measured by ACT and SAT mathematics percentile scores (ACTM%). Gender was accessed from university records where it was recorded as a binary variable. This work coded gender as a dichotomous variable with levels 0 (women) and 1 (men). This coding of gender is not optimal but is consistent with other work in PER. For a more nuanced discussion of gender, see Traxler *et al.* [266].

7.2.3 Mediation and Moderation

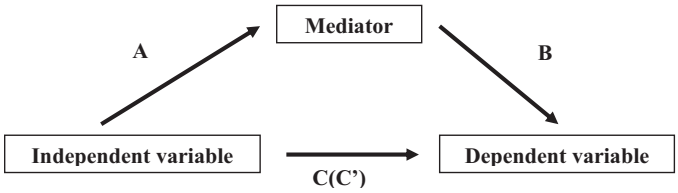


Figure 7.1: Mediation Process

Mediation and moderation form a powerful framework for investigating the relationships between variables affecting educational achievement. To investigate the relationship of personality facets and self-efficacy with achievement, the mediation framework developed by Baron and Kenny [120] was used. Figure 7.1 represents the mediational model for the relations of the dependent variable (Dep), independent variable ($Indep$), and the mediator (Med). The dependent variable of each regression is the node at the tail of the directed line; the independent variables for the regression are all nodes at the head of lines directed at the

dependent variable.

The total effect, C, is measured through the regression in Equation7.1,

$$Dep = \beta_1 + C \cdot Indep + \epsilon_1 \tag{7.1}$$

where β_i is the intercept and ϵ_i is the residual error.

With the mediator, Dep is predicted through two paths: the direct path characterized by C' and the indirect path through the mediator composed of a path from $Indep$ to Med (A) and the path from Med to Dep (B). These parameters are measured by Equations 7.2 and 7.3.

$$Med = \beta_2 + A \cdot Indep + \epsilon_2 \tag{7.2}$$

$$Dep = \beta_3 + C' \cdot Indep + B \cdot Med + \epsilon_3 \tag{7.3}$$

Significant mediation exists if A, B, and C are significant regression coefficients and if the direct effect C' is less than the total effect C. If $C' < C$, part of the overall effect of $Indep$ on Dep is a result of the relation of both variables with the mediator. To further test for significant mediation, the total indirect effect ($A \cdot B$) was calculated by bootstrapping with 1000 replications. The mediation is significant if a t-test shows this product is significantly different than zero. The the total effect of the independent variable on the dependent variable, C, is thus partitioned into two parts: one resulting from the mediator ($A \cdot B$) and one not resulting from the mediator (C'). The total effect can be expressed as a sum of these two contributions, $C = C' + A \cdot B$. The percentage of the total effect C which is the

result of the mediator is then $A \cdot B/C \cdot 100\%$.

Moderation occurs when one variable, the moderator (Mod), influences the relation between two other variables. For example, it may be that the relation of self-efficacy to course grade is different for men and women; in this case, gender may moderate the relationship between self-efficacy and course grade. To detect moderation, the moderator is added to the regression equation as a product term as shown in Equation7.4.

$$Dep = \beta_4 + \beta_5 \cdot Indep + D \cdot Indep \cdot Mod + \epsilon_4 \tag{7.4}$$

The moderation is significant if D is significant. If the moderator is dichotomous, the effect of moderation is to produce different slopes, β_5 and $\beta_5 + D$, depending on the level of the moderator (0 or 1, respectively).

7.3 Results

Table 7.1 shows the descriptive statistics for each course. The effect size difference between men and women for each variable is measured by Cohen's d. For each variable, a t-test was performed to determine if the difference between men and women was significant; the result of the t-test is presented as a superscript on d. A Bonferroni correction was applied to each of Tables 7.1 to 7.4 individually to correct for the inflation of Type I error. The significance threshold was divided by the number of statistical tests performed in the table. For example, in Table 7.1 the p threshold was divided by 36 for the 36 statistical tests performed in the columns from self-efficacy to openness which are the focus of this work. The grade and ACTM% columns are presented for reference and as a general measure of

Table 7.1: Descriptive statistics. Cohen's d measures the effect size for the difference between men and women for each quantity. The significance of a t-test of the difference between men and women is shown as a superscript on d. Note: "a" denotes $p < 0.05$, "b" $p < 0.01$, and "c" $p < 0.001$. A Bonferroni correction was applied to the significance levels. "M" refers to Men and "W" Women. "Gr" refers to Grade, "SEF" to Self-Efficacy, "Agr" to Agreeableness, "Cns" to Conscientiousness, "Ext" to Extraversion, "Nrt" to Neuroticism, and "Opn" to Openness.

	N	ACTM%	Gr	SEF	Agr	Cns	Ext	Nrt	Opn
Workshop Mathematics									
M	752	41.0 ± 12.5	2.9 ± 1.2	3.9 ± 0.8	3.8 ± 0.5	3.6 ± 0.6	3.2 ± 0.7	2.8 ± 0.7	3.5 ± 0.5
W	322	39.6 ± 11.7	2.9 ± 1.3	3.9 ± 0.9	3.9 ± 0.5	3.6 ± 0.6	3.3 ± 0.7	3.3 ± 0.8	3.4 ± 0.5
d		0.12	0.00	0.01	0.30^c	0.05	0.16	0.70^c	0.18
Calculus 1A									
M	387	71.2 ± 16.4	2.6 ± 1.2	3.9 ± 0.8	3.8 ± 0.6	3.6 ± 0.6	3.2 ± 0.7	2.7 ± 0.7	3.6 ± 0.5
W	378	67.6 ± 17.3	2.5 ± 1.3	3.7 ± 0.9	3.9 ± 0.6	3.7 ± 0.6	3.2 ± 0.8	3.3 ± 0.8	3.5 ± 0.6
d		0.21	0.05	0.27^b	0.18	0.18	0.04	0.69^c	0.05
Calculus 1B									
M	304	72.6 ± 16.8	2.7 ± 1.1	3.9 ± 0.8	3.7 ± 0.6	3.7 ± 0.6	3.2 ± 0.7	2.7 ± 0.7	3.6 ± 0.5
W	259	70.1 ± 16.4	2.8 ± 1.1	3.8 ± 0.9	3.9 ± 0.6	3.8 ± 0.6	3.2 ± 0.8	3.2 ± 0.7	3.5 ± 0.6
d		0.15	0.05	0.13	0.22	0.14	0.00	0.80^c	0.04
Calculus 1									
M	1020	83.8 ± 12.9	2.3 ± 1.3	3.9 ± 0.8	3.8 ± 0.6	3.6 ± 0.6	3.2 ± 0.8	2.7 ± 0.7	3.6 ± 0.5
W	555	82.2 ± 12.8	2.6 ± 1.2	3.6 ± 1.0	3.9 ± 0.6	3.7 ± 0.6	3.3 ± 0.8	3.2 ± 0.8	3.6 ± 0.6
d		0.12	0.21^c	0.28^c	0.11	0.19^a	0.10	0.73^c	0.07
Physics 1									
M	1284	80.1 ± 15.3	2.9 ± 1.0	3.9 ± 0.8	3.8 ± 0.6	3.7 ± 0.6	3.2 ± 0.7	2.6 ± 0.7	3.6 ± 0.5
W	499	81.3 ± 14.6	2.9 ± 1.0	3.6 ± 1.0	3.9 ± 0.6	3.9 ± 0.6	3.3 ± 0.8	3.1 ± 0.7	3.7 ± 0.6
d		0.08	0.01	0.36^c	0.24^c	0.32^c	0.07	0.70^c	0.08
Physics 2									
M	1186	81.5 ± 15.1	2.9 ± 1.0	3.9 ± 0.8	3.8 ± 0.6	3.7 ± 0.6	3.2 ± 0.7	2.7 ± 0.7	3.6 ± 0.5
W	365	83.5 ± 12.8	3.1 ± 1.0	3.8 ± 0.9	3.9 ± 0.6	3.8 ± 0.6	3.2 ± 0.7	3.1 ± 0.8	3.7 ± 0.6
d		0.14	0.15	0.23^a	0.20	0.19	0.07	0.64^c	0.01

differences in the student populations of each class.

The data presented have some striking features. In all courses, women report higher levels of neuroticism with effect sizes ranging from $d = 0.64$ to 0.80, from a medium to a large effect. The 0.4 to 0.6 Likert point difference on the neuroticism scale was very consistent between the physics and calculus classes. In addition, the Workshop Mathematics class, which is taken by a population of students with a lower percentage intending on majoring in

STEM and students with less mathematics preparation (as measured by ACTM%), shows a similar gender difference. As such, it seems likely that these differences represent a general feature of college-age students, not a specific feature of students pursuing physical science and engineering majors.

Women also consistently report higher levels of conscientiousness in calculus and physics classes with effect sizes for the differences ranging from below a small effect $d = 0.14$ to a small effect $d = 0.32$; however, these differences were significant only in Calculus 1 and Physics 1.

Men report higher levels of self-efficacy toward their calculus and physics classes with differences ranging from an effect size of 0.13 to 0.36 with the difference in the range of a small effect in all calculus and physics classes except Calculus 1B. The differences were statistically significant in all calculus and physics classes except Calculus 1B. Self-efficacy toward the remedial Workshop Mathematics class was approximately equal for men and women. Men reported the same level of self-efficacy as that reported in the more challenging calculus and physics classes. Women in Workshop Mathematics reported higher levels of self-efficacy than women in the calculus and physics classes. This may have resulted from the class being fairly easy with simply completing assignments all that was required for a passing grade.

Overall, the averages of the personality facets of men and women were strikingly similar to classes requiring substantially different high school preparation and bridging the first two years of college. Significant differences between men and women were also measured for the agreeableness facet in Workshop Mathematics and Physics 1 with women reporting higher levels of this facet.

The sample contains two two-class course sequences: Calculus 1A and 1B and Physics 1 and 2. For both sequences, the difference between the self-efficacy of men and women was smaller for the second class in the course sequence. The self-efficacy of women increased between the first class in the sequence and the second class. This change could be the result of increased self-efficacy in women with longer exposure to physics and mathematics. It could also be caused by women with lower self-efficacy choosing not to enroll in the second class in the sequence.

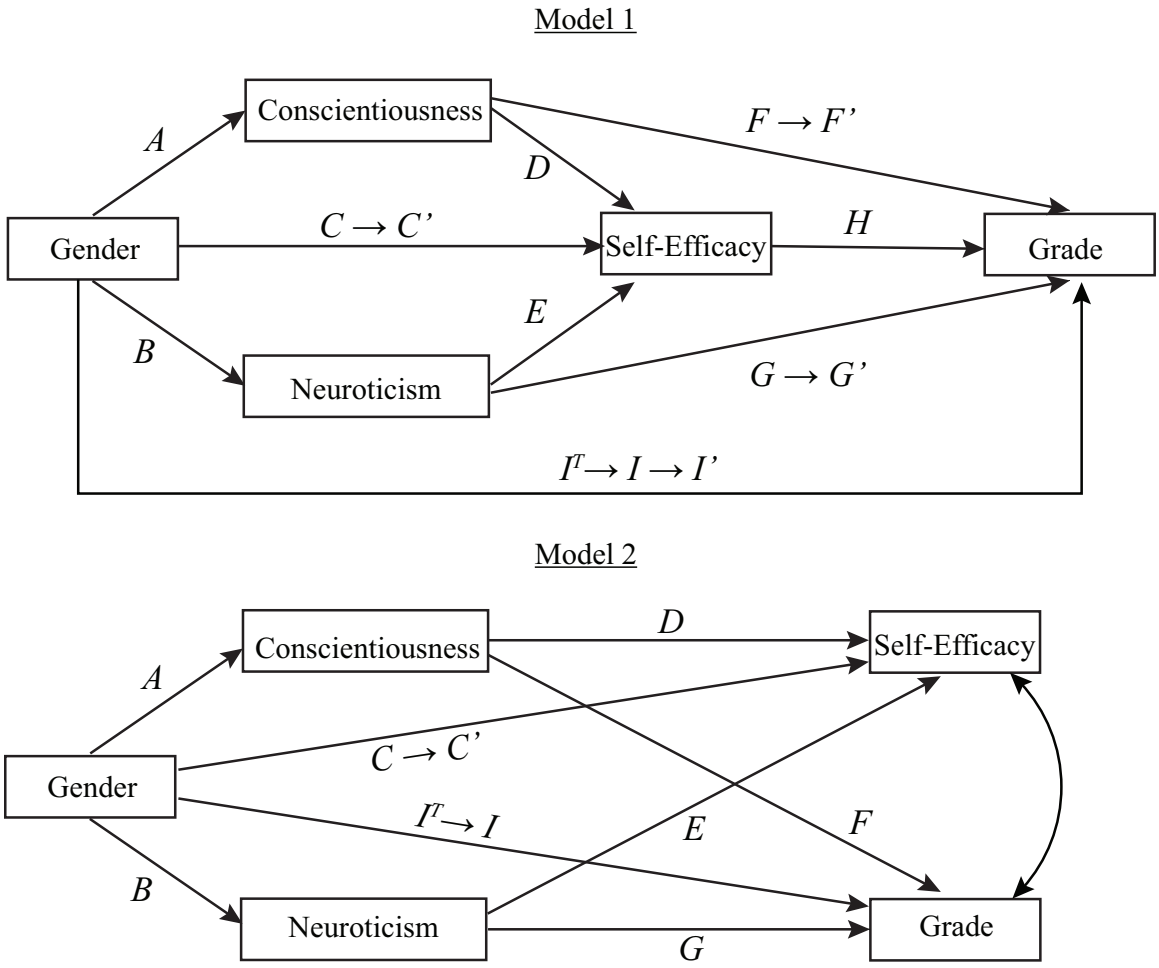


Figure 7.2: Path model showing the relation of gender, personality, self-efficacy, and physics course grade.

To further investigate this effect, a paired sample was extracted consisting of students

who had taken both courses in the course sequence. For Physics 1 and Physics 2, 865 students took both courses (men 635, women 230); the self-efficacy of women did not significantly increase from Physics 1 (3.7 ± 1.0) to Physics 2 (3.8 ± 0.9), the effect size of this difference is $d = 0.24$, nor did the self-efficacy of men, Physics 1 (4.0 ± 0.8) to Physics 2 (4.0 ± 0.8), effect size $d = 0.10$. A similar matched sample was extracted for the Calculus 1A and 1B sequence ($N = 312$, men 157, women 155). The self-efficacy of women did not significantly increase from Calculus 1A (3.8 ± 0.9) to Calculus 1B (3.8 ± 0.9), $d = 0.08$; the self-efficacy of men also did not significantly change from Calculus 1A (4.0 ± 0.7) to Calculus 1B (3.9 ± 0.8), $d = 0.21$. This result was similar to the finding of Cwik and Singh's work, reporting a consistent self-efficacy at the beginning and at the end of the course [179].

7.3.1 Full Path Model

Figure 7.2 shows two possible path models for the relation of gender, the personality facets conscientiousness and neuroticism, self-efficacy, and physics course grade. Only conscientiousness and neuroticism were examined because of the significant differences observed in Table 7.1, prior work relating these variables to academic achievement, and for the theoretical reasons discussed in Sec. 7.1.3. Model 2 contains an additional element; the curved line between self-efficacy and grade represents the correlation between these variables. Treated as structural equation models (SEM), these path models are mathematically equivalent. In SEM, reversing the direction of an edge or replacing an edge with a correlation does not change the overall fit of the model. Two models are presented to allow the investigation of different assumptions of the relation of self-efficacy to grade. A path model encodes the relational hypotheses of the researcher.

In both models, our relational hypothesis is that gender influences personality which in turn influences self-efficacy. This hypothesis is supported by the consistent gender difference in the neuroticism facet observed in both STEM and non-STEM samples; that gender identity usually develops prior to adult personality; and that personality develops generally prior to STEM self-efficacy. It is also supported by causal arguments relating higher anxiety or conscientiousness to academic performance which informs the development of self-efficacy.

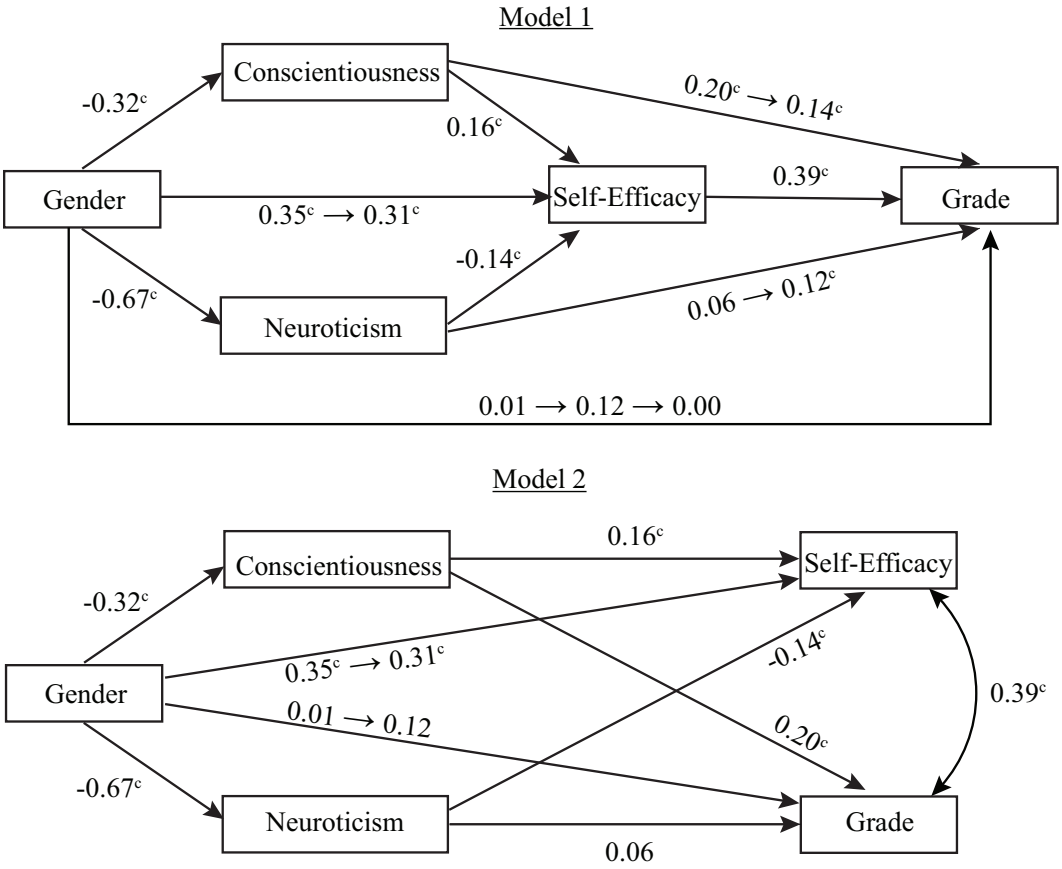


Figure 7.3: Path models showing the relation of gender, personality, and self-efficacy for students in Physics 1. Gender was coded with women as zero, and men as one. The number on each path is the value of the regression coefficient. The notation #1 → #2 shows the change in the coefficient before (#1) and after (#2) the addition of the mediating variables. Compare the figure with Figure7.2 for the symbolic variable related to each number. Note: "a" denotes $p < 0.05$, "b" $p < 0.01$, and "c" $p < 0.001$. A Bonferroni correction was applied to the significance levels.

For Model 1, we additionally assume self-efficacy influences grades, but that personality also influences grades directly and through its effect on self-efficacy. This model is supported

theoretically by the generally positive effect of believing one can do something on actually doing that thing. It is also supported temporally; self-efficacy is measured mid-semester while grades are assigned at the end of the semester.

For Model 2, we discard the hypothesis that changes in self-efficacy imply changes in grades and relax it to the assumption that self-efficacy covaries with grades. This model can be theoretically justified by observing self-efficacy is related to prior course success which should influence grades. Modifying self-efficacy should influence future grades; however, modifying self-efficacy will not modify prior academic performance which also is related to course grades.

In Model 1 or 2, personality could affect self-efficacy either by modifying how past experience is processed into current beliefs, by modifying the past experiences themselves, or by doing both. The effect of personality on self-efficacy through either mechanism has been acting for many years; as such, the structure in Models 1 and 2 relating gender, personality, and self-efficacy should be viewed as a cumulative effect acting over many years.

The central difference between Models 1 and 2 is that in Model 1 changes in self-efficacy should produce changes in course grades while in Model 2 differences in self-efficacy should be related to differences in grades. There are two possible interpretations of Model 1. The first views both self-efficacy and grade as quantities measured at a single instance in time and suggests that if some intervention could improve self-efficacy prior to the end of the course then this change would have an effect on the current course grade (as well as future course grades) suggested by the coefficient of the path model. We argue that this interpretation is unlikely (particularly given the size of the measured coefficients). An intervention to modify self-efficacy at a given time would not change the past experiences (academic achievement)

that informed self-efficacy. A second interpretation acknowledges the recursive nature of Bandura's model and views both self-efficacy and academic achievement (measured by a college course grade) as variables that have developed over time. In this view, personality has influenced self-efficacy over the student's development which has in turn influenced their general academic achievement.

This work investigates three potential mediational relationships shown in Model 1; two of these relations are also present in Model 2. The first relation explores the mediation of the combination of conscientiousness and neuroticism of the relationship between gender and self-efficacy. This mediation model is composed of edges A, B, $C \rightarrow C'$, D, and E. The notation $C \rightarrow C'$ indicates that the coefficient C representing the total effect of gender on self-efficacy is changed to C' by the action of the mediating variables. The second mediational relation investigates whether the relation of gender to grade is mediated by personality. This model is composed of edges: A, B, F, G, and $I^T \rightarrow I$. The third possible mediating relationship shown only in Model 1 investigates whether self-efficacy mediates the relation of gender and personality to grade; this model requires the full path Model 1. These three analyses will be discussed in the next three sections.

The path model in Figure7.2 was analyzed with traditional multiple linear regression analysis. It could also be analyzed as a structural equation model (SEM) which yields the same results (it is a saturated or just-identified model so all model fit statistics are perfect). For the SEM model, the two personality facets are assumed to co-vary.

The path models for Physics 1 are shown in Figure7.3. The discussion which follows focuses on Physics 1 when a specific example is needed, but also discusses the general features of all classes.

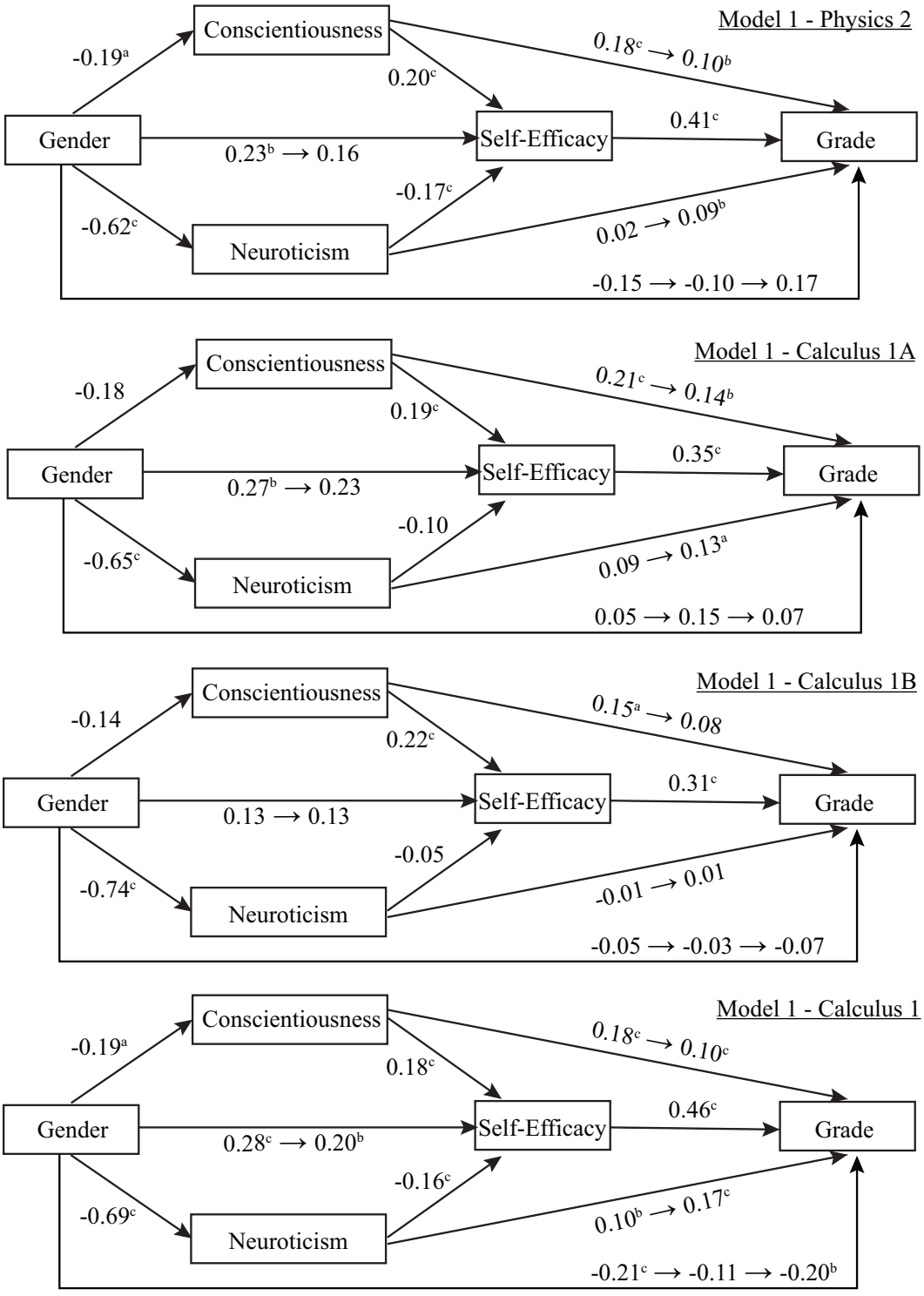


Figure 7.4: Path models showing the relation of gender, personality, and self-efficacy for students in Physics 2, Calculus 1A, Calculus 1B, and Calculus 1. The number on each path is the value of the regression coefficient. The notation #1 → #2 shows the change in the coefficient before (#1) and after (#2) the addition of the mediating variables. Compare the figure with Fig. 7.2 for the symbolic variable related to each number. Note: "a" denotes $p < 0.05$, "b" $p < 0.01$, and "c" $p < 0.001$.

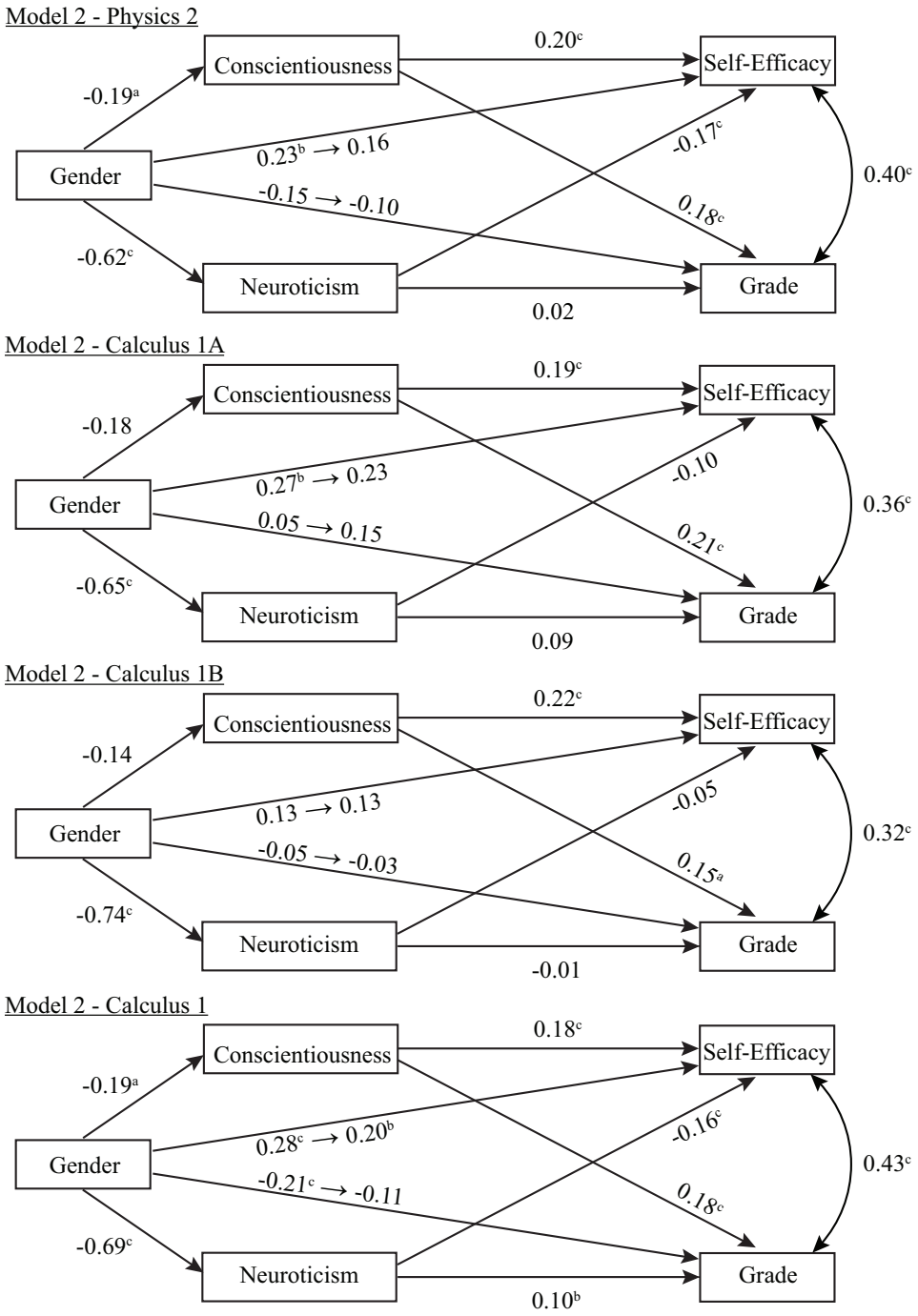


Figure 7.5: Model 2 path models showing the relation of gender, personality, and self-efficacy for students in Physics 2, Calculus 1A, Calculus 1B, and Calculus 1. The number on each path is the value of the regression coefficient. The notation #1 → #2 shows the change in the coefficient before (#1) and after (#2) the addition of the mediating variables. Compare the figure with Fig. 7.2 for the symbolic variable related to each number. Note: "a" denotes $p < 0.05$, "b" $p < 0.01$, and "c" $p < 0.001$.

Sections 7.3.2, 7.3.3, and 7.3.4 present three separate mediation analyses. The overall results of these analyses for Model 1 and Model 2 are presented in the path model in Fig. 7.3 for Physics 1. The path models presenting Model 1 for the other classes are shown in Figure7.4; the path models for Model 2 are presented in Figure7.5. Physics 1 is presented in the same figure to allow a comparison between the two models. Results will be discussed for all classes; when a specific example would be helpful, Physics 1 is used.

7.3.2 Mediation of the relation of gender to self-efficacy

Table 7.2: The mediation by neuroticism and conscientiousness of the relation of gender to self-efficacy. The regression coefficient β and its standard error (SE) are presented. Women are coded as zero, and men as one. For indirect effects, the product of the path coefficients $\beta_i\beta_j$ is presented and the standard deviation (SD) of the product. Conscientiousness is abbreviated Cns, neuroticism Nrt, and self-efficacy SEF. Note: "a" denotes $p < 0.05$, "b" $p < 0.01$, and "c" $p < 0.001$. A Bonferroni correction was applied to the significance levels. "Cal" refers to Calculus in "Cal 1A, Cal 1B and, Cal 1 and "Phys" refers to Physics in "Phys 1" and "Phys 2".

	Cal 1A		Cal 1B		Cal 1		Phys 1		Phys 2	
	β	SE	β	SE	β	SE	β	SE	β	SE
Total Effect and Remaining Effect										
Gender $\rightarrow$ SEF (C)	0.27^b	0.07	0.13	0.08	0.28^c	0.05	0.35^c	0.05	0.23^b	0.06
$C = C' + A \cdot D + B \cdot E$										
Gen $\rightarrow$ SE (C')	0.23	0.08	0.13	0.09	0.20^b	0.05	0.31^c	0.05	0.16	0.06
Direct Effects										
Gen $\rightarrow$ Cns (A)	-0.18	0.07	-0.14	0.08	-0.19^a	0.05	-0.32^c	0.05	-0.19^a	0.06
Gen $\rightarrow$ Nrt (B)	-0.65^c	0.05	-0.74^c	0.08	-0.69^c	0.05	-0.67^c	0.05	-0.62^c	0.06
Cns $\rightarrow$ SEF (D)	0.19^c	0.05	0.22^c	0.04	0.18^c	0.03	0.16^c	0.02	0.20^c	0.03
Nrt $\rightarrow$ SEF (E)	-0.10	0.04	-0.05	0.05	-0.16^c	0.03	-0.14^c	0.03	-0.17^c	0.03
Indirect Effects										
	$\beta_i\beta_j$	SD	$\beta_i\beta_j$	SD	$\beta_i\beta_j$	SD	$\beta_i\beta_j$	SD	$\beta_i\beta_j$	SD
Gen $\rightarrow$ Cns $\rightarrow$ SEF ($A \cdot D$)	-0.03^c	0.01	-0.03^c	0.02	-0.03^c	0.01	-0.05^c	0.01	-0.05^c	0.01
% of Total effect ($A \cdot D/C$)	-12.9%		-22.9%		-12.0%		-14.4%		-16.4%	
Gen $\rightarrow$ Nrt $\rightarrow$ SEF ($B \cdot E$)	0.07^c	0.03	0.04^c	0.03	0.11^c	0.02	0.09^c	0.01	0.11^c	0.01
% of Total effect ($B \cdot E/C$)	24.9%		26.6%		39.0%		26.7%		46.5%	

Differences in personality may affect how a student interacts with academic situations. Higher conscientiousness may lead a student to complete more of his or her assignments or to work harder on those assignments, thus increasing academic achievement. Higher neuroticism may add to the anxiety felt during testing, lowering academic achievement, or may cause a student to feel excess concern about the class causing an increase in effort and increasing achievement. Either lower or higher past academic achievement should influence future self-efficacy in Bandura's model. As such, it is possible personality mediates the relationship of gender and self-efficacy.

The mediation analysis for all physics and calculus courses is shown in Table 7.2. For all analyses that follow, all continuous variables have been normalized by subtracting the mean of each variable and dividing by the standard deviation. The regression coefficients were computed using Equation7.5 to 7.8 where SEF is self-efficacy, Nrt is neuroticism, Cns is conscientiousness, β_i^0 is the intercept and ϵ_i the residual error.

$$SEF = \beta_1^0 + C \cdot Gender + \epsilon_1 \tag{7.5}$$

$$Cns = \beta_2^0 + A \cdot Gender + \epsilon_2 \tag{7.6}$$

$$Nrt = \beta_3^0 + B \cdot Gender + \epsilon_3 \tag{7.7}$$

$$SEF = \beta_4^0 + C' \cdot Gender + D \cdot Cns + E \cdot Nrt + \epsilon_4 \tag{7.8}$$

The beta coefficient for the linear relation between a dichotomous and a normalized continuous variable represents the difference in the average of the continuous variable in standard deviation units between the two levels of the dichotomous variable and is related

to Cohen's d. The linear relation between two normalized continuous variables is related to the correlation between the two variables.

The relation of gender to self-efficacy was significant for all classes except Calculus 1B. For all four of these classes, the total effect of gender on self-efficacy (C) was reduced; however, in many classes, the amount of reduction was fairly small. Examination of the two indirect paths from gender to self-efficacy shows that this was the result of paths through neuroticism and conscientiousness partially cancelling with men having lower conscientiousness which led to lower self-efficacy (14.4% of the total effect for Physics 1) while the lower level of neuroticism in men led to higher self-efficacy (26.7% of the total effect in Physics 1). For all classes, the indirect path through neuroticism accounted for a higher percentage of the total effect of gender on self-efficacy than the path through conscientiousness. For three of the classes, the path through neuroticism accounted for about 25% of the total effect; for Calculus 1 and Physics 2, 40% of the total effect. As such, a substantial part of the gender difference in self-efficacy was explained by differences in neuroticism. The path through conscientiousness explained about 15% of the total effect.

The direct effect of gender on personality shows the same pattern as was observed in Table 7.1. The gender difference in conscientiousness was generally significant but had less than a small effect in all classes except Physics 1. The direct effect of gender on neuroticism was much larger, in the range of a medium effect. The pattern of direct effects was different for Calculus 1A and 1B and the other classes. Note, for dichotomous variables regression coefficient β is related to Cohen's d but they are not identical; d normalizes the difference in level by the pooled standard deviation while β uses the aggregated standard deviation.

7.3.3 Mediation of the relation of gender to achievement

For this section, we consider the overall effect of gender on course grade (I^T) which is estimated by Equation7.9 and whether the personality facets conscientiousness and neuroticism mediate this relationship. This mediation model is formed of the edges A, B, F, G, and $I^T \rightarrow I$ in Figure7.1 Model 2; these edges also appear in Model 1 with F, G, and I the values of the regression coefficients before taking into account the mediation of self-efficacy.

$$Grade = \beta_7^0 + I^T \cdot Gender + \epsilon_7 \tag{7.9}$$

The coefficients F, G, and I are estimated by Equation7.10.

$$Grade = \beta_5^0 + F \cdot Cns + G \cdot Nrt + I \cdot Gender + \epsilon_5 \tag{7.10}$$

Table 7.3 summarizes the mediation analysis. The total effect of gender on grade, I^T, was significant only in Calculus 1 at the level of a small effect. The effect of the personality variables either reduced the female advantage in grades or increased the male advantage. Examination of the indirect effects showed that the reason for this change is that women gain a small advantage in course grades through both the indirect path through conscientiousness and neuroticism. The advantage through the path through conscientiousness was expected and is supported by many general education studies [167]. The advantage gained through the path through neuroticism was less expected, but still understandable. Any disadvantage accrued through higher neuroticism causing increased anxiety in testing situations must be offset by the positive impacts of feeling anxiety or other strong emotions. For example,

Table 7.3: The mediation by neuroticism and conscientiousness of the relation of gender to grade. The regression coefficient β and its standard error (SE) are presented. Women are coded as zero, and men as one. For indirect effects, the product of the path coefficients $\beta_i\beta_j$ is presented and the standard deviation (SD) of the product. Conscientiousness is abbreviated Cns, neuroticism Nrt, and self-efficacy SEF. Note: "a" denotes $p < 0.05$, "b" $p < 0.01$, and "c" $p < 0.001$. A Bonferroni correction was applied to the significance levels. "Cal" refers to Calculus in "Cal 1A, Cal 1B and, Cal 1 and "Phys" refers to Physics in "Phys 1" and "Phys 2". "Gen" refers to Gender, "Gr" to Grade, "SEF" to Self-Efficacy, "Agr" to Agreeableness, "Cns" to Conscientiousness, "Ext" to Extraversion, "Nrt" to Neuroticism, and "Opn" to Openness.

	Cal 1A		Cal 1B		Cal 1		Phys 1		Phys 2	
	β	SE	β	SE	β	SE	β	SE	β	SE
Total Effect and Remaining Effect										
Gen→Gr (I^T)	0.05	0.08	-0.05	0.08	-0.21^c	0.05	0.01	0.05	-0.15	0.06
$I^T = A \cdot F + B \cdot G + I$										
Gen → Gr (I)	0.15	0.08	-0.03	0.09	-0.11	0.06	0.12	0.06	-0.10	0.06
Direct Effects										
Gen → Cns (A)	-0.18	0.07	-0.14	0.08	-0.19^a	0.05	-0.32^c	0.05	-0.19^a	0.06
Gen → Nrt (B)	-0.65^c	0.05	-0.74^c	0.08	-0.69^c	0.05	-0.67^c	0.05	-0.62^c	0.06
Nrt → Gr (G)	0.09	0.04	-0.01	0.05	0.10^b	0.05	0.06	0.03	0.02	0.03
Cns → Gr (F)	0.21^c	0.04	0.15^a	0.04	0.18^c	0.03	0.20^c	0.02	0.18^c	0.03
Indirect Effects										
	$\beta_i\beta_j$	SD	$\beta_i\beta_j$	SD	$\beta_i\beta_j$	SD	$\beta_i\beta_j$	SD	$\beta_i\beta_j$	SD
Gen → Cns → Gr ($A \cdot F$)	-0.04^c	0.02	-0.02^c	0.02	-0.04^c	0.01	-0.06^c	0.01	-0.03^c	0.01
Gen → Nrt → Gr ($B \cdot G$)	-0.06^c	0.02	0.01	0.04	-0.06^c	0.02	-0.04^c	0.02	-0.01^c	0.02

additional anxiety prior to a test may cause a student to prepare more thoroughly for the test.

The mediation analysis in Tables 7.2 and 7.3 can be combined into Model 2 in Fig. 7.2. All coefficients have been presented except the correlation between self-efficacy and grade. This correlation is presented in Table 7.4 and is fairly large (a medium effect in all classes).

7.3.4 Mediation of the relation of personality and self-efficacy to achievement

Substantial research has demonstrated a relationship between personality and academic achievement [167]. The previous sections demonstrated that personality, particularly the neuroticism facet, mediated gender differences in self-efficacy and that the personality facets

	Cal 1A		Cal 1B		Cal 1		Phys 1		Phys 2	
	β	SE	β	SE	β	SE	β	SE	β	SE
Total and Remaining Effects										
Gen $\rightarrow$ Gr (I)	0.15	0.08	−0.03	0.09	−0.11	0.06	0.12	0.06	−0.10	0.06
$I = C' \cdot H + I'$										
Gen $\rightarrow$ Gr (I')	0.07	0.07	−0.07	0.09	-0.20^{b}	0.05	0.00	0.05	−0.17	0.06
Cns $\rightarrow$ Gr (F)	0.21^{c}	0.04	0.15^{a}	0.04	0.18^{c}	0.03	0.20^{c}	0.02	0.18^{c}	0.03
$F = D \cdot H + F'$										
Cns $\rightarrow$ Gr (F')	0.14^{b}	0.04	0.08	0.04	0.10^{c}	0.02	0.14^{c}	0.02	0.10^{b}	0.02
Nrt $\rightarrow$ Gr (G)	0.09	0.04	−0.01	0.05	0.10^{b}	0.05	0.06	0.03	0.02	0.03
$G = E \cdot H + G'$										
Nrt $\rightarrow$ Gr (G')	0.13^{a}	0.04	0.01	0.04	0.17^{c}	0.02	0.12^{c}	0.02	0.09^{b}	0.03
Direct Effects										
SEF $\rightarrow$ Gr (H)	0.35^{c}	0.03	0.31^{c}	0.04	0.46^{c}	0.02	0.39^{c}	0.02	0.41^{c}	0.02
Indirect Effects										
	$\beta_i\beta_j$	SD	$\beta_i\beta_j$	SD	$\beta_i\beta_j$	SD	$\beta_i\beta_j$	SD	$\beta_i\beta_j$	SD
Gen $\rightarrow$ SEF $\rightarrow$ Gr										
($C' \cdot H$)	0.09^{c}	0.02	0.03^{c}	0.02	0.10^{c}	0.03	0.12^{c}	0.02	0.12^{c}	0.02
% of Total effect										
($C' \cdot H/I$)	–		–		–		–		–	
Cns $\rightarrow$ SEF $\rightarrow$ Gr										
($D \cdot H$)	0.06^{c}	0.01	0.07^{c}	0.02	0.08^{c}	0.01	0.06^{c}	0.01	0.06^{c}	0.01
% of Total effect										
($D \cdot H/F$)	32.1%		45.8%		44.7%		30.5%		44.1%	
Nrt $\rightarrow$ SEF $\rightarrow$ Gr										
($E \cdot H$)	-0.03^{c}	0.01	-0.01^{c}	0.01	-0.07^{c}	0.01	-0.06^{c}	0.01	-0.05^{c}	0.01
% of Total effect										
($E \cdot H/G$)	–		–		-71.6%		–		–	
Correlations										
	r		r		r		r		r	
Cor. SEF and Gr	0.36^{c}		0.32^{c}		0.43^{c}		0.39^{c}		0.40^{c}	

Table 7.4: The mediation by self-efficacy of the relation of gender, neuroticism, and conscientiousness to grade. The regression coefficient β and its standard error (SE) are presented. Women are coded as zero, and men as one. For indirect effects, the product of the path coefficients $\beta_i\beta_j$ is presented and the standard deviation (SD) of the product. Conscientiousness is abbreviated Cns, neuroticism Nrt, and self-efficacy SEF. Note: "a" denotes $p < 0.05$, "b" $p < 0.01$, and "c" $p < 0.001$. A Bonferroni correction was applied to the significance levels. "Cal" refers to Calculus in "Cal 1A, Cal 1B and, Cal 1 and "Phys" refers to Physics in "Phys 1" and "Phys 2". "Gen" refers to Gender, "Gr" to Grade, "SEF" to Self-Efficacy, "Agr" to Agreeableness, "Cns" to Conscientiousness, "Ext" to Extraversion, "Nrt" to Neuroticism, and "Opn" to Openness.

modified the relation of gender to grade. Self-efficacy has been reliably demonstrated as one of the most important non-cognitive factors in explaining academic achievement [167]. It is, therefore, possible that the reported relation between personality and academic achievement actually exist because personality affects self-efficacy which affected achievement.

The unmediated model for this analysis removes the self-efficacy node from the path model in Fig. 7.2 - Model 1 and was investigated in the previous section. It contains the edges A estimated by Equation7.6, B estimated by Equation7.7, and F, G, and I estimated by Equation7.10.

The full model in Figure7.2 Model 1 forms the mediated model containing self-efficacy. The addition of self-efficacy potentially modified the effect of conscientiousness on grade changing F to F', neuroticism on grade changing G to G', and the effect of gender on grade changing I to I'. These coefficients as well as the direct effect of self-efficacy on grade (H) are estimated by Equation7.11.

$$\begin{aligned} Grade = \beta_6^0 + F' \cdot Cns + G' \cdot Nrt + H \cdot SEF + \\ I' \cdot Gender + \epsilon_6 \end{aligned} \tag{7.11}$$

Each of the total effects can be partitioned into a remaining direct effect and an effect through the mediator (SEF): $F = F' + D \cdot H$, $G = G' + E \cdot H$, and $I = I' + C' \cdot H$. The fraction of the total effect that acts through the mediator is then calculated. The results for all classes are shown in Table7.4 and the full path model for Physics 1 in Figure7.3 Model 1.

The total effect of gender on grade controlling for personality (I) was small and not

significant in all classes after a Bonferroni correction was applied. Self-efficacy modified this effect (I') exposing a significant advantage toward women in course grade in Calculus 1 controlling for personality and self-efficacy. In all classes, the mediated effect of gender on grade was more advantageous to women than the unmediated effect.

The total effect of conscientiousness on grade controlling for gender and neuroticism (F) was significant and at or near the level of a small effect in all classes. This was consistent with a substantial body of research showing the importance of this facet in explaining academic performance [167]. This effect was strongly mediated by self-efficacy (F') with the path through self-efficacy accounting for 30% to 45% of the total effect of conscientiousness on grade. As such, a substantial portion of this facet's effect on academic performance can be explained by its effect on self-efficacy. This is consistent with Bandura's model where the prior academic achievement experienced by conscientious students leads to higher levels of self-efficacy.

The total effect of neuroticism on grade (G) controlling for conscientiousness and gender was small in all classes and significant at the $p < 0.01$ level in only one class. The addition of self-efficacy exposed a significant positive effect of neuroticism on grades in four of the five classes; the effect was at or near the level of a small effect.

7.3.5 Moderation

The models of the previous section were further analyzed to determine if gender moderated the relations in the models.

Gender was added as a moderator to Equations7.8, 7.9, and 7.10 to determine if the effect of personality was different for men and women. For example, Equation7.8 was modified

to Equation7.12.

$$\begin{aligned} SEF = \beta_8^0 + & C' \cdot Gender + D \cdot Cns + \\ & M_D \cdot Gender \cdot Cns + \\ & E \cdot Nrt + M_E \cdot Gender \cdot Nrt + \epsilon_8 \end{aligned} \tag{7.12}$$

where M_D and M_E are the regression coefficients of the interaction terms (moderators). If either regression coefficient is significant, the relation of either conscientiousness or neuroticism to self-efficacy is different for men and women. For example, if M_D is significant, the slope of the relation between conscientiousness and self-efficacy is D for women and $D + M_D$ for men.

No statistically significant moderation was found in any of the courses (the regression coefficients M_D and M_E were not significant in any course). This is especially important and shows that, while women report higher mean levels of conscientiousness and neuroticism than men, the relation of both to self-efficacy is the same for men and women.

The moderation of the relation of conscientiousness, neuroticism, and self-efficacy to grade was tested with Equation7.13.

$$\begin{aligned} Grade = \beta_9^0 + & I' \cdot Gender + F' \cdot Cns + \\ & M_F \cdot Gender \cdot Cns + \\ & G \cdot Nrt + M_G \cdot Gender \cdot Nrt + \\ & H \cdot SEF + M_H \cdot Gender \cdot SEF + \epsilon_9 \end{aligned} \tag{7.13}$$

where M_F, M_G, and M_H are the regression coefficients of the interaction terms. No statis-

tically significant moderation was found in any of the courses.

7.4 Discussion

This study was conducted to investigate five research questions. Each question will be discussed in order in this section.

RQ1: Does self-efficacy or personality differ for men and women in core university introductory mathematics and physics classes? The self-efficacy of men and women were significantly different in Calculus 1A, Calculus 1, Physics 1, and Physics 2. The results for these courses were consistent with Huang's 2013 meta-analysis which showed the self-efficacy of men was higher than women in STEM classes [233]. The effect size for the difference in self-efficacy for Physics 1 was reduced in Physics 2, similarly, the effect size of the difference for Calculus 1A was reduced in Calculus 1B and was no longer significant.

Multiple personality facets were different for men and women in some classes. Agreeableness was significantly different in Workshop Mathematics and Physics 1 with women reporting higher levels of agreeableness in each of these courses; the differences represented a small effect. Women reported significantly higher conscientiousness in Calculus 1 and Physics 1, a small effect. Neither openness nor extraversion were significantly different for men and women in any class.

Women reported significantly higher neuroticism in all classes, a medium to large effect. These values differed by 0.4 to 0.6 points on a 5-point Likert scale. The observed differences in neuroticism were similar to those reported in a large ($N > 10^6$) non-academic study [267]. The consistency of the differences in this facet for all classes and compared to a national

sample suggest the difference in neuroticism is a common feature of college-age students, not a feature specific to STEM students.

RQ2: Does personality mediate the relationship of gender to self-efficacy? If so, how does it mediate the relationship? The mediation of self-efficacy by the personality facets conscientiousness and neuroticism was investigated in Table7.2. For Calculus 1A, Calculus 1, Physics 1, and Physics 2 the total effect (C) of gender on self-efficacy was significant (a small effect) and fairly similar with β ranging from 0.23 to 0.35. With neuroticism and conscientiousness added as mediating variables, these total effects were reduced somewhat by 0.04 to 0.07 to produce remaining direct effects for 0.16 to 0.31. Examination of the path model showed this weak mediation partially resulted from the competition of the two facets. The lower conscientiousness of men led to lower self-efficacy accounting for an average of 16% of the total effect (C). The lower neuroticism of men led to higher self-efficacy accounting for an average of 33% of the total effect.

RQ3: Does personality mediate the relationship of gender to achievement? If so, how does it mediate the relationship? This model is summarized in Table7.3. The relation of gender to achievement was significant in only one class so in general the relation fails Baron and Kenny's test of mediation. The direct effect of gender on conscientiousness was significant in three of the five classes at or near the level of a small effect. We note the coefficients for Calculus 1A and 1B are also near a small effect and the failure to find a significant effect in these classes is likely the result of the smaller sample size. The coefficient was significant in Calculus 1A before the Bonferroni correction. The direct effect of gender on neuroticism was significant and large in all classes at the level of medium to large effect. The direct effect of conscientiousness on grades was significant and positive in all classes at the level of a small

effect. The direct effect of neuroticism on grades was generally positive but significant in only one class. The indirect effects through both conscientiousness and neuroticism were significant and negative in all classes except Calculus 1B. The sum of these effects was -0.10 in three of the classes; half the size of a small effect. This indicates in general that women have an advantage in achievement both due to higher levels of conscientiousness and neuroticism, reflecting a small effect in the first class in each sequence (Calculus 1A, Calculus 1, and Physics 1).

RQ4: Does self-efficacy mediate the relationship of personality and gender to achievement? If so, how does it mediate the relationship? Self-efficacy could potentially mediate the relation of three variables to achievement in this model; the total effect of gender on grade (I), the total effect of conscientiousness on grade (F), and the total effect of neuroticism on grade (G). The effect of gender on grade was small and non-significant in all classes; the addition of self-efficacy exposed a significant (small effect size) advantage to women, but only in Calculus 1. Conscientiousness had a significant positive total effect on grade (F) in all classes, a small effect. The β coefficients were strongly reduced in all classes and became insignificant in Calculus 1B; self-efficacy strongly mediated the relation of conscientiousness to grade explaining on average 39% of the total effect. As such, a substantial part of the effect of conscientiousness on grades can be explained by its prior effect on self-efficacy. The total effect of neuroticism on grade (G) was small in all classes; it was significant at the level of a small effect only in Calculus 1. With the addition of self-efficacy as a mediator, a significant remaining positive direct effect (G') was uncovered in four or five classes at or near the level of a small effect. As such, higher neuroticism improves grades once the negative effect of self-efficacy is accounted for. Conscientiousness and neuroticism produced

competing indirect effects on grades by their action on self-efficacy.

RQ5: Does gender moderate the relationships of personality, self-efficacy, and achievement? No significant moderation was detected in any model. The relationship of conscientiousness and neuroticism to self-efficacy is the same for men and women, as is the relationship of conscientiousness, neuroticism, and self-efficacy to achievement.

7.5 Implications and recommendations

This work combined multiple well-established research strands: the relation of anxiety to test performance, the relation of achievement to self-efficacy, general differences between how men and women report the tendency to experience anxiety, general differences in conscientiousness between men and women, and the relation of conscientiousness to academic success. Together, these strands suggest that a substantial amount of the often-reported differences in self-efficacy between men and women may result from competing gender differences in the tendency to experience anxiety and the tendency to conscientiously complete tasks. Self-efficacy has long been an important construct in models explaining career choice and persistence and is a a significant contributor to academic success. As such, variations in physics and mathematics self-efficacy by gender may be one source of differences in the representation of men and women in STEM fields requiring these classes.

This work advanced a more nuanced definition of self-efficacy for students in college science and mathematics classes. For students just starting their journey in the sciences, a belief they can succeed may be separate from experiences informing that belief. Interventions that change self-efficacy can act on those beliefs without the confounding variable of prior

success. Students in the mathematics and physics classes have long experience with mathematics and science classes and generally a history of success in those classes. Interventions to modify self-efficacy may change beliefs, but cannot modify prior experiences upon which those beliefs are grounded. As such, care should be taken in interpreting the relationship between self-efficacy and achievement as causal as implied in Figure7.2 - Model 1 as opposed to correlational as shown in Model 2. An intervention increasing self-efficacy will likely not increase grade to the extent implied in Model 1 because a substantial part of the relation of self-efficacy to grade must result from the relation of prior achievement to grade which informs self-efficacy but also affects grades.

The work presented suggests that some modification of self-efficacy is needed. Model 2 in Figure7.2 shows that while the higher conscientiousness of women is related to higher grades as well as higher self-efficacy as Bandura's model suggests should be the case, the higher neuroticism of women which was also related to higher grades was related to lower self-efficacy.

Having identified differences in personality differences affect achievement and self-efficacy, and identified substantial prior academic experiences as an important component of the academic self-efficacy of college STEM students, one can re-examine interventions designed to improve self-efficacy.

7.6 Limitations

This work was performed at a single research university in calculus and calculus-based physics classes. Additional research at other institutions including primarily teaching-focused

institutions with different student populations is needed to understand if the results obtained in this research can be generalized to represent physics students nationally. Additional research should also investigate algebra-based physics classes. Further, this study used a single observation of self-efficacy collected mid-semester. Multiple measurements were taken at different times during the semester and longitudinally in different classes would allow a more thorough characterization of the recursive development of self-efficacy predicted by Bandura's model. The study also captures generally self-efficacy toward the class; this self-efficacy could be differentiated between differing tasks within the class.

7.7 Conclusions and Future Research

This study identified differences in personality as a potential origin for the differences in the self-efficacy beliefs of men and women in physics and mathematics courses. Personality may also explain differences often reported for men and women in engineering classes; most of the students in the current study were engineering majors. Similar differences in self-efficacy are not reported in chemistry and biology while the students in these classes almost certainly also have the same differences in conscientiousness and neuroticism reported in this study. Future research should be conducted to understand the features of course environments that both promote and constrain the development of students' physics and mathematics self-efficacy beliefs. Beyond these possible directions, a qualitative study could shed further light on the self-efficacy difference between men and women in physics classes.

This work examined the conscientiousness and neuroticism facets of the five-factor model of personality and self-efficacy toward physics and mathematics for students in intro-

ductory physics and mathematics classes. Women reported substantially higher neuroticism in all courses studied, near a large effect. This was consistent with the results of a large national study suggesting the result is general. Women also reported higher conscientiousness and lower self-efficacy in many of the classes studied, with small effects. Neuroticism mediated the relation of gender to self-efficacy substantially in most classes; the path through the mediator explained from 25% to 47% of the total effect. Conscientiousness mediated the effect of gender on self-efficacy more weakly explaining for 12% to 23% of the total effect.

The relation of personality to self-efficacy and self-efficacy to course grade was generally consistent for men and women; significant moderation was not measured in any class. As such, the negative relation of neuroticism to self-efficacy is the same for men and women.

Chapter 8

Future Work

Although this work explored a number of academic and non-cognitive factors that affect students' college physics achievement, not every factor that can influence college physics achievement was investigated in this study. Some potential future and ongoing projects are outlined below:

- Researchers interested in investigating the impact of high school physics preparation should collect more comprehensive data regarding the high school experience including in-class pedagogy. This will enable the exploration of a wider range of factors that influence on high school physics preparation.

- This study demonstrated that both pretest and post-test scores exhibited relations with overall high school achievement, as measured by ACT/SAT scores. These relations have the potential to skew normalized gain in favor of populations with higher scores on these assessments, as normalized gain tends to favor such populations. Alternative statistical measures that can assess conceptual gain should be developed.

- The analysis of AP courses should be extended to other enriched curricula, such as the International Baccalaureate (IB) program. Collaborative research involving multiple institutions could help address the scarcity of participants, enabling a more comprehensive exploration of the factors influencing achievement in college physics.

The work discussed in Chapters 4, 5 and 6 indicated that the CLASS instrument is described by a complex model, not completely explained by a simple factor model. This study lays the foundation for future research to uncover the underlying complexities and gain a more comprehensive understanding of the relationships between the variables in the research domain.

The work discussed in Chapter 7 suggests that men and women process their prior academic achievement differently because they feel different levels of anxiety. This result has implications for designing, developing, and modifying instructional structures to reduce anxiety.

Chapter 9

Conclusion

The development of modern PER applying DBER methodologies has allowed physics educators to investigate the factors that affect students' college achievements; multiple studies have reported the relation of non-cognitive factors on college achievement [18, 19]. Chapter 3 investigated factors influencing Force and Motion Conceptual Evaluation (FMCE) pretest and post-test scores in the introductory calculus-based mechanics class at a large eastern land-grant university.

Several academic and non-cognitive factors were examined using correlation analysis and linear regression analysis to understand their relation to students' physics conceptual understanding. Students' prior exposure to high school physics impacted their performance in the FMCE and the type of high school physics, whether it was regular physics or AP physics, dramatically changed the impact. Pretest scores completely captured the effect of high school preparation on post-test scores.

Chapters 4, 5 and 6 investigated the factor structures suggested by Adam *et al.*, [1] and Douglas *et al.*, [2] for the CLASS using correlation analysis, EFA and CFA. In the current study, both EFA and CFA suggested that the initial eight-factor model introduced by Adams *et al.* did not fit the instrument well; the three-factor model introduced by Douglas *et al.* was an improvement but also failed to describe the instrument well. A four subscale model was developed using the Douglas *et al.* three-factor model as the basis. The model was then tuned to produce good fit model parameters. The final tuning suggested that the factors were not orthogonal and that multiple items and subscales were correlated with each other.

Chapter 7 discussed how the personality facets conscientiousness and neuroticism as well as self-efficacy were related to physics and mathematics for students enrolled in introductory physics and mathematics courses. Women reported significantly higher levels of

neuroticism across all the courses analyzed, with an effect size approaching a large effect. Women reported higher conscientiousness and lower self-efficacy in several of the courses considered, with small effect sizes. In most classes, neuroticism played a substantial mediating role in the relationship between gender and self-efficacy, while conscientiousness had a weaker mediating effect.

In general, the investigation of non-cognitive factors in physics provided an additional valuable tool for understanding physics success beyond traditional measures of academic achievement.

Bibliography

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

[2] K.A. Douglas, M.S. Yale, D.E. Bennett, M.P. Haugan, and L.A. Bryan. Evaluation of Colorado Learning Attitudes about Science Survey. *Phys. Rev. Spec. Topics-Phys. Educ. Res.*, 10(2):020128, 2014.

[3] M. Estrada, M. Burnett, A.G. Campbell, P.B. Campbell, W.F. Denetclaw, C.G. Gutiérrez, S. Hurtado, G.H. John, J. Matsui, R. McGee, et al. Improving underrepresented minority student persistence in STEM. *CBE—Life Sci. Educ.*, 15(3), 2016.

[4] AAU Undergraduate STEM Education Initiative. A five-year status report on the AAU undergraduate STEM education initiative. *Association of American Universities*, 2017.

[5] E. Mach. *The Science of Mechanics: A Critical and Historical Exposition of its Principles*. Open Court Publishing Company, Chicago, IL, 1893.

[6] P.W. Bridgman. *The Logic of Modern Physics*, volume 3. Macmillan, New York, NY, 1927.

[7] S. Devons and L. Hartmann. A history-of-physics laboratory. *Phys. Today*, 23(2):44–49, 1970.

[8] D. Baird, R.I. Hughes, and A. Nordmann. *Heinrich Hertz: Classical Physicist, Modern Philosopher*, volume 198. Springer Science & Business Media, New York, NY, 1998.

[9] P.A.M. Dirac. The relation between mathematics and physics. *Proceedings of the Roy. Soc. Edinburgh*, 59:122–129, 1940.

[10] H.T. Hudson and W.R. McIntire. Correlation between mathematical skills and success in physics. *Am. J. Phys.*, 45(5):470–471, 1977.

[11] A.B. Champagne, L.E. Kopfer, and J.H. Anderson. Factors influencing learning of classical mechanics. *Proceedings of the Am. Educ. Res. Asso.*, 1979.

[12] I.A. Halloun and D. Hestenes. The initial knowledge state of college physics students. *Am. J. Phys.*, 53(11):1043, 1985.

[13] J.H. Larkin and G.C. Brackett. Mathematics pre-requisites: A mastery approach. *Am. J. Phys.*, 42(12):1089–1091, 1974.

[14] A.B. Champagne and L.E. Klopfer. A causal model of students' achievement in a college physics course. *J. Res. Sci. Teach.*, 19(4):299–309, 1982.

[15] G.J. Pallrand and F. Seeber. Spatial ability and achievement in introductory physics. *J. Res. Sci. Teach.*, 21(5):507–516, 1984.

[16] L. Leboutet-Barrell. Concepts of mechanics among young people. *Phys. Educ.*, 11(7):462, 1976.

[17] J.W. Renner. Significant physics content and intellectual development—cognitive development as a result of interacting with physics content. *Am. J. Phys.*, 44(3):218–222, 1976.

[18] S.R. Singer, N.R. Nielsen, and H.A. Schweingruber. *Discipline-Based Education Ressearch*. National Academies Press, Washington, DC, 2012.

[19] J. L. Docktor and J. P. Mestre. Synthesis of discipline-based education research in physics. *Phys. Rev. Phys. Educ. Res.*, 10(2):020119, 2014.

[20] K. Cummings. A developmental history of physics education research. In *Second Committee Meeting on the Status, Contributions, and Future Directions of Discipline-Based Education Research.*, Alexandria, VA, 2011. The National Science Foundation.

[21] D. Hestenes, M. Wells, and G. Swackhamer. Force concept inventory. *Phys. Teach.*, 30(3):141–158, 1992.

[22] R.R. Hake. Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *Am. J. Phys.*, 66:64–74, 1998.

[23] L.C. McDermott and E.F. Redish. Resource letter: PER-1: Physics education research. *Am. J. Phys.*, 67(9):755–767, 1999.

[24] R.K. Thornton. Conceptual dynamics: Following changing student views of force and motion. In *AIP Conf. Proc.*, volume 399, College Park, MD, 1997. American Institute of Physics.

[25] E. Bagno and B.S. Eylon. From problem solving to a knowledge structure: An example from the domain of electromagnetism. *Am. J. Phys.*, 65(8):726–736, 1997.

[26] M.G.M. Ferguson-Hessler and T. de Jong. Studying physics texts: Differences in study processes between good and poor performers. *Cognition Instruct.*, 7(1):41–54, 1990.

[27] E. Cohen, A. Mason, C. Singh, and E. Yerushalmi. Identifying differences in diagnostic skills between physics students: Students' self-diagnostic performance given alternative scaffolding. In *AIP Conf. Proc.*, volume 1064, College Park, MD, 2008. American Institute of Physics.

[28] E. Yerushalmi, A. Mason, E. Cohen, and C. Singh. Effect of self diagnosis on subsequent problem solving performance. In *AIP Conf. Proc.*, volume 1064, College Park, MD, 2008. American Institute of Physics.

[29] B. Ibrahim and N.S. Rebello. Representational task formats and problem solving strategies in kinematics and work. *Phys. Rev. Spec. Topics-Phys. Educ. Res.*, 8(1):010126, 2012.

[30] D. Rosengrant, A. Van Heuvelen, and E. Etkina. Free-body diagrams: Necessary or sufficient? In *AIP Conf. Proc.*, volume 790, College Park, MD, 2005. American Institute of Physics.

[31] A. Van Heuvelen and X. Zou. Multiple representations of work–energy processes. *Am. J. Phys.*, 69(2):184–194, 2001.

[32] E.T. Torigoe and G.E. Gladding. Connecting symbolic difficulties with failure in physics. *Am. J. Phys.*, 79(1):133–140, 2011.

[33] B.L. Sherin. How students understand physics equations. *Cognition Instruct.*, 19(4):479–541, 2001.

[34] L. Cui, N.S. Rebello, and A.G. Bennett. College students' transfer from calculus to physics. In *AIP Conf. Proc.*, volume 818, College Park, MD, 2006. American Institute of Physics.

[35] E. Cohen and S.E. Kanim. Factors influencing the algebra "reversal error". *Am. J. Phys.*, 73(11):1072–1078, 2005.

[36] P. Heller and M. Hollabaugh. Teaching problem solving through cooperative grouping. part 2: Designing problems and structuring groups. *Am. J. Phys.*, 60(7):637–644, 1992.

[37] J.P. Mestre. Probing adults' conceptual understanding and transfer of learning via problem posing. *J. Appl. Dev. Psychol.*, 23(1):9–50, 2002.

[38] E.F. Redish. *Teaching Physics with the Physics Suite*. American Association of Physics Teachers, College Park, MD, 2004.

[39] A.B. Arons and T.D. Miner. *A Guide to Introductory Physics Teaching*. American Association of Physics Teachers, College Park, MD, 1990.

[40] R.D. Knight. *Five Easy Lessons: Strategies for Successful Physics Teaching*. American Association of Physics Teachers, College Park, MD, 2004.

[41] D.E. Meltzer and R.K. Thornton. Resource letter ALIP–1: Active-learning instruction in physics. *Am. J. Phys.*, 80(6):478–496, 2012.

[42] 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(1):64–74, 1998.

[43] D.E. Meltzer and K. Manivannan. Transforming the lecture-hall environment: The fully interactive physics lecture. *Am. J. Phys.*, 70(6):639–654, 2002.

[44] T. De Jong and M.G. Ferguson-Hessler. Cognitive structures of good and poor novice problem solvers in physics. *J. Educ. Psychol.*, 78(4):279, 1986.

[45] F. Reif and J.I. Heller. Knowledge structure and problem-solving in physics. *Educ. Psychol.*, 17(2):102–127, 1982.

[46] E.F. Redish. A theoretical framework for physics education research: Modeling student thinking. *arXiv preprint physics/0411149*, 2004.

[47] H. Helm and J.D Novak. *Proceedings of the International Seminar Misconceptions in Science and Mathematics, June 20-22, 1983*. Cornell University, New York, NY, 1983.

[48] J. Confrey. Chapter 1: A review of the research on student conceptions in mathematics, science, and programming. *Rev. Res. Educ.*, 16(1):3–56, 1990.

[49] S. Flores-Garcia, L.L. Alfaro-Avena, J.E. Chavez-Pierce, J. Luna-Gonzalez, and M.D. Gonzalez-Quezada. Students' difficulties with tension in massless strings. *Am. J. Phys.*, 78(12):1412–1420, 2010.

[50] R.A. Streveler, T.A. Litzinger, R.L. Miller, and P.S. Steif. Learning conceptual knowledge in the engineering sciences: Overview and future research directions. *J. Eng. Educ.*, 97(3):279–294, 2008.

[51] J. Clement. Students' preconceptions in introductory mechanics. *Am. J. Phys.*, 50(1):66–71, 1982.

[52] G.J. Posner, K.A. Strike, P.W. Hewson, and W.A. Gertzog. Toward a theory of conceptual change. *Sci. Educ.*, 66(2):211–227, 1982.

[53] J.D. Bransford, A.L. Brown, R.R. Cocking, et al. *How People Learn*, volume 11. National Academy Press, Washington, DC, 2000.

[54] A. Gupta, D. Hammer, and E.F. Redish. The case for dynamic models of learners' ontologies in physics. *J. Learn. Sci.*, 19(3):285–321, 2010.

[55] C.W. Camp, J.J. Clement, D. Brown, K. Gonzalez, J. Kudukey, J. Minstrell, K. Schultz, M. Steinberg, V. Veneman, and A. Zietsman. *Preconceptions in Mechanics: Lessons Dealing with Students' Conceptual Difficulties*. Citeseer, Princeton, NJ, 2010.

[56] L.C. McDermott and P.S. Shaffer. *Tutorials in Introductory Physics*. Prentice Hall, Upper Saddle River, NJ, 2002.

[57] J.D. Slotta and M.T.H. Chi. Helping students understand challenging topics in science through ontology training. *Cognition Instruct.*, 24(2):261–289, 2006.

[58] Paula V Engelhardt. *Getting Started in PER*, volume 2. AAPT, College Park, MD, 2009.

[59] R.S. Lindell, E. Peak, and T.M. Foster. Are they all created equal? a comparison of different concept inventory development methodologies. In *AIP Conf. Proc.*, volume 883, pages 14–17, College Park MD, 2007. American Institute of Physics.

[60] D. Hestenes and M. Wells. A mechanics baseline test. *Phys. Teach.*, 30(3):159–166, 1992.

[61] 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(S1):S12, 2001.

[62] 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(4):338, 1998.

[63] R. K. Thornton, D. Kuhl, K. Cummings, and J. Marx. Comparing the Force and Motion Conceptual Evaluation and the Force Concept Inventory. *Phys. Rev. Phys. Educ. Res.*, 5(1):010105, 2009.

[64] D. Hewagallage, E. Christman, and J. Stewart. Examining the relation of high school preparation and college achievement to conceptual understanding. *Phys. Rev. Spec. Topics-Phys. Educ. Res.*, 18(1):010149, 2022.

[65] I. Halloun, R.R. Hake, E.P. Mosca, and D. Hestenes. Force Concept Inventory (revised 1995), 1995. `http://modeling.asu.edu/R&E/Research.html` Accessed 7/19/2019.

[66] D.B. May and E. Etkina. College physics students' epistemological self-reflection and its relationship to conceptual learning. *Am. J. Phys.*, 70(12):1249–1258, 2002.

[67] K.K. Perkins, W.K. Adams, S.J. Pollock, N.D. Finkelstein, and C.E. Wieman. Correlating student beliefs with student learning using the colorado learning attitudes about science survey. In *AIP Conf. Proc.*, volume 790, College Park, MD, 2005. American Institute of Physics.

[68] D. Hammer. Epistemological beliefs in introductory physics. *Cog. Ins.*, 12(2):151–183, 1994.

[69] E. Yerushalmi, C. Henderson, K. Heller, P. Heller, and V. Kuo. Physics faculty beliefs and values about the teaching and learning of problem-solving. *Phys. Rev. Spec. Topics-Phys. Educ. Res.*, 3(2):020109, 2007.

[70] C. Henderson and M.H. Dancy. Barriers to the use of research-based instructional strategies: The influence of both individual and situational characteristics. *Phys. Rev. Spec. Topics-Phys. Educ. Res.*, 3(2):020102, 2007.

[71] C. Henderson and M.H. Dancy. Physics faculty and educational researchers: Divergent expectations as barriers to the diffusion of innovations. *Am. J. Phys.*, 76(1):79–91, 2008.

[72] C. Turpen and N.D. Finkelstein. The construction of different classroom norms during peer instruction: Students perceive differences. *Phys. Rev. Spec. Topics-Phys. Educ. Res.*, 6(2):020123, 2010.

[73] C. Turpen and N.D. Finkelstein. Not all interactive engagement is the same: Variations in physics professors' implementation of peer instruction. *Phys. Rev. Spec. Topics-Phys.s Educ. Res.*, 5(2):020101, 2009.

[74] E.F. Redish, J.M. Saul, and R.N. Steinberg. Student expectations in introductory physics. *Am. J. Phys.*, 66(3):212–224, 1998.

[75] I. Halloun. Views about Science and Physics Achievement: The VASS Story. In *AIP Conf. Proc.*, volume 399, pages 605–614, College Park, MD, 1997. American Institute of Physics.

[76] S.D. Willoughby and K. Johnson. Epistemic beliefs of non-STEM majors regarding the nature of science: Where they are and what we can do. *Am. J. Phys.*, 85(6):461–468, 2017.

[77] 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. Phys. Educ. Res.*, 9:020121, Nov 2013.

[78] B. Toven-Lindsey, M. Levis-Fitzgerald, P.H. Barber, and T. Hasson. Increasing persistence in undergraduate science majors: A model for institutional support of underrepresented students. *CBE-Life Sci. Educ.*, 14(2):1–12, 2015.

[79] P. Zhang, L. Ding, and E. Mazur. Peer instruction in introductory physics: A method to bring about positive changes in students' attitudes and beliefs. *Phys. Rev. Phys. Educ. Res.*, 13(1):010104, 2017.

[80] V.K. Otero and K.E. Gray. Attitudinal gains across multiple universities using the physics and everyday thinking curriculum. *Phys. Rev. ST Phys. Educ. Res.*, 4(2):020104, 2008.

[81] E. Brewe, A. Traxler, J. de la Garza, and L.H. Kramer. Extending positive class results across multiple instructors and multiple classes of modeling instruction. *Phys. Rev. ST Phys. Educ. Res.*, 9(2):020116, 2013.

[82] P. Zhang and L. Ding. Large-scale survey of chinese precollege students' epistemological beliefs about physics: A progression or a regression? *Phys. Rev. ST Phys. Educ. Res.*, 9(1):010110, 2013.

[83] K. Gray, W. Adams, C. Wieman, and K. Perkins. Students know what physicists believe, but they don't agree: A study using the CLASS survey. *Phys. Rev. ST Phys. Educ. Res.*, 4(2):020106, November 2008.

[84] G.J. Privitera. *Essential Statistics for the Behavioral Sciences*. Sage Publications, Los Angeles, CA, 2017.

[85] R. Likert. A technique for the measurement of attitudes. *Arch. Psychol.*, 140:5–55, 1932.

[86] J. Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Academic Press, New York, NY, 2013.

[87] K. Kelley and K.J. Preacher. On effect size. *Psychol. Method.*, 17(2):137, 2012.

[88] C.J. Ferguson. An effect size primer: a guide for clinicians and researchers. *Am. Psychol. Assoc.*, 2016.

[89] C. Ialongo. Understanding the effect size and its measures. *Biochemia medica*, 26(2):150–163, 2016.

[90] N.J Gogtay and U.M. Thatte. Principles of correlation analysis. *J. Asso. Physic. India*, 65(3):78–81, 2017.

[91] K.U Gülden and G. Neşe. A study on multiple linear regression analysis. *Procedia - Social and Behavioral Sciences*, 106:234–240, 2013.

[92] P.T. Pope and J.T. Webster. The use of an F-statistic in stepwise regression procedures. *Technometrics*, 14(2):327–340, 1972.

[93] J.P.C. Kleijnen. Cross-validation using the t statistic. *Eur. J. Oper. Res.*, 13(2):133–141, 1983.

[94] C. Spearman. General intelligence, objectively determined and measured, 1904. *Am. J. Phys.*, 100(3):697, 1987.

[95] C.R. Rao and S. Sinharay. *Handbook of Statistics 26: Psychometrics*. North Holland, Amsterdam, Netherland, 2006.

[96] L.R. Fabrigar and D.T. Wegener. *Exploratory Factor Analysis*. Oxford University Press, Northamptonshire, UK, 2011.

[97] B. Williams, A. Onsman, and T. Brown. Exploratory factor Analysis: A five-step guide for novices. *Aust. J. Paramedicine*, 8(3), 2010.

[98] L.R. Fabrigar, D.T. Wegener, R.C. MacCallum, and E.J. Strahan. Evaluating the use of exploratory factor analysis in psychological research. *Psychol. Method.*, 4(3):272, 1999.

[99] T.A. Brown and M.T. Moore. *Handbook of Structural Equation Modeling*, volume 361. The Guilford Press, New York, NY, 2012.

[100] M.A. Pett, N.R. Lackey, J.J. Sullivan, and *et. al. Making Sense of Factor Analysis: The Use of Factor Analysis for Instrument Development in Health Care Research*. Sage Publications, Thousand Oaks, CA, 2003.

[101] B. Thompson. Exploratory and Confirmatory Factor Analysis: Understanding concepts and applications. *Appl. Psychol. Meas.*, 10694(000):3, 2004.

[102] K.Y. Hogarty, C.V. Hines, J.D. Kromrey, J.M. Ferron, and K.R. Mumford. The quality of factor solutions in exploratory factor analysis: The influence of sample size, communality, and overdetermination. *Educ. and Psychol. Measurement*, 65(2), 2005.

[103] B.G. Tabachnick, L.S. Fidell, and J.B. Ullman. *Using Multivariate Statistics*, volume 5. Pearson, Boston, MA, 2007.

[104] A.L. Comrey and H.B. Lee. *A First Course in Factor Analysis*. Psychology Press, London, UK, 2013.

[105] K.G. Sapnas and R.A. Zeller. Minimizing sample size when using exploratory factor analysis for measurement. *J. Nurs. Meas.*, 10(2), 2002.

[106] R.K. Henson and J.K. Roberts. Use of exploratory factor analysis in published research: Common errors and some comment on improved practice. *Educ. and Psychol. Meas.*, 66(3), 2006.

[107] H.F. Kaiser. The application of electronic computers to factor analysis. *Educ. Psychol. Meas.*, 20(1), 1960.

[108] R.B. Cattell. The scree test for the number of factors. *Multivar. Behav. Res.*, 1(2), 1966.

[109] J.L. Horn. A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2), 1965.

[110] E.E. Cureton and R.B. D'Agostino. *Factor Analysis: An Applied Approach*. Psychology Press, London, UK, 2013.

[111] R.L. Gorsuch. *Factor Analysis: Classic Edition*. Routledge, New York, NY, 2014.

[112] R.E. Schumacker and R.G. Lomax. *A Beginner's Guide to Structural Equation Modeling*. Psychology Press, United Kingdom, 2004.

[113] R.B. Kline. *Principles and Practice of Structural Equation Modeling*. Guilford Publications, New York, NY, 2015.

[114] J.B. Ullman and P.M. Bentler. *Handbook of Psychology, Second Edition*, volume 2. Wiley Online Library, 2012.

[115] R.C. Dahiya and J. Gurland. Pearson chi-squared test of fit with random intervals. *Biometrika*, 59(1):147–153, 1972.

[116] P.M. Bentler. Comparative fit indexes in structural models. *Psychol. Bull.*, 107(2):238, 1990.

[117] H.W. Marsh, J.R. Balla, and R.P. McDonald. Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychol. Bull.*, 103(3):391, 1988.

[118] D.A. Kenny, B. Kaniskan, and D.B. McCoach. The performance of RMSEA in models with small degrees of freedom. *Socio. Method. Res.*, 44(3):486–507, 2015.

[119] G. Taasoobshirazi and S. Wang. The performance of the SRMR, RMSEA, CFI, and TLI: An examination of sample size, path size, and degrees of freedom. *J. Appl. Quant. Method.*, 11(3):31–39, 2016.

[120] R.M. Baron and D.A. Kenny. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *J. Abnorm. Psychol. Soc. Psychol.*, 51(6):1173, 1986.

[121] D.P. MacKinnon, A.J. Fairchild, and M.S. Fritz. Mediation analysis. *Annu. Rev. Psychol.*, 58:593, 2007.

[122] D.A. Kenny. Reflections on mediation. *Organ. Res. Method.*, 11(2):353–358, 2008.

[123] B.K. Prahani, I. Limatahu, S.W. Winata, L. Yuanita, and M. Nur. Effectiveness of physics learning material through guided inquiry model to improve student's problem-solving skills based on multiple representations. *Int. J. Educ. Res.*, 4(12):231–244, 2016.

[124] D.M. Dimitrov and Phillip D. Rumrill Jr. Pretest-posttest designs and measurement of change. *Work*, 20(2):159–165, 2003.

[125] A. Suyatna, D. Anggraini, D. Agustina, and D. Widyastuti. The role of visual representation in physics learning: dynamic versus static visualization. In *J. Phys.: Conference Series*, volume 909, page 012048, Bristol, England, 2017. IOP Publishing.

[126] 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(2):020114, 2019.

[127] 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(1):010107, 2021.

[128] 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(12):1259, 2002.

[129] 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(1):010131, 2019.

[130] J. Yang, C. Zabriskie, and J. Stewart. Multidimensional item response theory and the Force and Motion Conceptual Evaluation. *Phys. Rev. Phys. Educ. Res.*, 15(2):020141, 2019.

[131] J.M. Nissen and J.T. Shemwell. Gender, experience, and self-efficacy in introductory physics. *Phys. Rev. Phys. Educ. Res.*, 12(2):020105, 2016.

[132] R. Henderson, D. Hewagallage, J. Follmer, L. Michaluk, J. Deshler, E. Fuller, and J. Stewart. Mediating role of personality in the relation of gender to self-efficacy in physics and mathematics. *Phys. Rev. Phys. Educ. Res.*, 18(1):010143, 2022.

[133] J. Aalst. An introduction to physics education research. *Can. J. Phys.*, 78(1):57–71, 2000.

[134] L.B. Sheeber, E.D. Sorensen, and S.R. Howe. Data analytic techniques for treatment outcome studies with pretest/posttest measurements: an extensive primer. *J. Psych. Res.*, 30(3):185–199, 1996.

[135] I.A. Halloun and D. Hestenes. Common sense concepts about motion. *Am. J. Phys.*, 53(11):1056, 1985.

[136] A. Savinainen and P. Scott. Using the Force Concept Inventory to monitor student learning and to plan teaching. *Phys. Educ.*, 37(1):53, 2002.

[137] A. Savinainen and P. Scott. The Force Concept Inventory: A tool for monitoring student learning. *Phys. Educ.*, 37(1):45, 2002.

[138] M.D. Caballero, E.F. Greco, E.R. Murray, K.R. Bujak, M. Jackson Marr, R. Catrambone, M.A. Kohlmyer, and M.F. Schatz. Comparing large lecture mechanics curricula using the Force Concept Inventory: A five thousand student study. *Am. J. Phys.*, 80(7):638–644, 2012.

[139] J. Docktor and K. Heller. Gender differences in both Force Concept Inventory and introductory physics performance. *AIP Conf. Proc.*, 1064:15–18, 2008.

[140] R. Chabay and B. Sherwood. Qualitative understanding and retention. *AAPT Announcer*, 27:96, 1997.

[141] Physport. `https://www.physport.org`. Accessed 8/8/2017.

[142] 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(4):245, 2017.

[143] S. Freeman, S.L. Eddy, M. McDonough, M.K. Smith, N. Okoroafor, H. Jordt, and M.Pat. Wenderoth. Active learning increases student performance in science, engineering, and mathematics. *P. Nat. Acad. Sci. USA*, 111(23):8410–8415, 2014.

[144] C.M. Schroeder, T.P. Scott, T.Y. Tolson, H.and 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(10):1436, 2007.

[145] P.L. Bonate. *Analysis of Pretest-Posttest Designs*. CRC Press, Boca Raton, FL, 2000.

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

[147] 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(1):010115, 2018.

[148] D.S. Hewagallage, J. Stewart, and R. Henderson. Differences in the Predictive Power of Pretest Scores of Students Underrepresented in Physics. In *Physics Education Research Conference 2020*, PER Conference, Washington, DC, 2020.

[149] D. Liberman and H.T. Hudson. Correlation between logical abilities and success in physics. *Am. J. Phys.*, 47(9):784, 1979.

[150] 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(12):1117, 1982.

[151] A.B. Champagne, L.E. Klopfer, and J.H. Anderson. Factors influencing the learning of classical mechanics. *Am. J. Phys.*, 48(12):1074, 1980.

[152] 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(1):010106, 2007.

[153] 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, 2002. Submitted to the Physics Education Research Conference, Boise, ID.

[154] G.E. Hart and P.D. Cottle. Academic backgrounds and achievement in college physics. *Phys. Teach.*, 31(8):470, 1993.

[155] 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(6):847–876, 2007.

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

[157] 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(1):23, 2015.

[158] 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(6):373, 2004.

[159] K.A. Koenig, M.C. Frey, and D.K. Detterman. ACT and general cognitive ability. *Intel.*, 36(2):153, 2008.

[160] 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. *Coll. Board Res. Pap.*, 2019.

[161] Jeff Allen. Updating the ACT college readiness benchmarks. ACT Research Report Series 2013 (6). *ACT, Inc.*, 2013.

[162] W.J. Camara and G. Echternacht. The SAT I and high school grades: Utility in predicting success in college. *Res. Not.*, 10, 2000.

[163] S.l 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):1, 2002.

[164] 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. *Ctr. Stud. High. Educ.*, 2007.

[165] W.G. Bowen, M.M. Chingos, and M.S. McPherson. *Measuring Success: Testing, Grades, and the Future of College Admissions*. Johns Hopkins University Press, Baltimore, MD, 2018.

[166] 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(6):2077, 2019.

[167] M. Richardson, C. Abraham, and R. Bond. Psychological correlates of university students' academic performance: A systematic review and meta-analysis. *Psychol. Bull.*, 138(2):353, 2012.

[168] A.E. Poropat. A meta-analysis of the five-factor model of personality and academic performance. *Psychol. Bull.*, 135(2):322, 2009.

[169] A. Bandura. Self-efficacy: Toward a unifying theory of behavioral change. *Psychol. Rev.*, 84(2):191, 1977.

[170] S. Andrew. Self-efficacy as a predictor of academic performance in science. *J. Adv. Nurs.*, 27(3):596, 1998.

[171] S. Lau and R.W. Roeser. Cognitive abilities and motivational processes in high school students' situational engagement and achievement in science. *Educ. Assessment*, 8(2):139, 2002.

[172] W.J. Hughes. Perceived gender interaction and course confidence among undergraduate science, mathematics, and technology majors. *J. Women Minor. Sci. Engineer.*, 6(2), 2000.

[173] 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(4):391, 2018.

[174] 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(2):020106, 2016.

[175] M.E. Junge and B.J. Dretzke. Mathematical self-efficacy gender differences in gifted/talented adolescents. *Gifted Child Quar.*, 39(1):22, 1995.

[176] 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. Eng. Educ.*, 90(4):477, 2001.

[177] 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. *Phy. Rev. Phys. Educ. Res.*, 12(2):020124, 2016.

[178] V. Sawtelle, E. Brewe, and L.H. Kramer. Positive impacts of modeling instruction on self-efficacy. In *AIP Conf. Proc.*, volume 1289, page 289, College Park, MD, 2010. American Institute of Physics.

[179] 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, Nov 2021.

[180] L.R. Goldberg. The development of markers for the big-five factor structure. *Psychol. Assessment*, 4(1):26, 1992.

[181] O.P. John, E.M. Donahue, and R.L. Kentle. Big Five Inventory. *J. Pers. Soc. Psychol.*, 1991.

[182] O.P. John, L.P. Naumann, and C.J. Soto. *Handbook of Personality: Theory and Research*. The Guilford Press, New York, NY, 2008.

[183] J.C. Stewart and D. Hewagallage. The relation of personality, gender, and achievement in science classes. In *AERA Annual Meeting*, Washington, DC, 2019.

[184] 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(1):103, 2010.

[185] X. Ma. A meta-analysis of the relationship between anxiety toward mathematics and achievement in mathematics. *J. Res. Math. Educ.*, 30:520–540, 1999.

[186] J.V. Mallow and S.L. Greenburg. Science anxiety: Causes and remedies. *J. Coll. Sci. Teach.*, 11:356–358, 1982.

[187] 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(4):435–446, 2004.

[188] 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(4):356–369, 2010.

[189] 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(3-4):175–183, 2012.

[190] S. Ramlo. Validity and reliability of the Force and Motion Conceptual Evaluation. *Am. J. Phys.*, 76(9):882, 2008.

[191] 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, April 2020.

[192] 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(2):020103, 2018.

[193] 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(2):020102, 2014.

[194] 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(2):020101, 2008.

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

[196] 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(1):010103, 2018.

[197] US News & World Report: Education. US News and World Report, Washington, DC. `https://premium.usnews.com/best-colleges`. Accessed 4/30/2017.

[198] 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. *J. Res. Personality*, 44(3):315, 2010.

[199] 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.

[200] T.G. Duncan and W.J. McKeachie. The making of the Motivated Strategies for Learning Questionnaire. *Educ. Psych.*, 40(2):117, 2005.

[201] 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(3):337, 2007.

[202] 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(4):700, 2012.

[203] R Core Team, 2017. https://www.R-project.org/Accessed on 12/01/2023.

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

[205] S. Epskamp, A.O.J. Cramer, J.L. Waldorp, V.D. Schmittmann, and D. Borsboom. qgraph: Network visualizations of relationships in psychometric data. *J. Stat. Soft.*, 2012.

[206] W.K. Adams, C.E. Wieman, K.K. Perkins, and J. Barbera. Modifying and validating the Colorado Learning Attitudes about Science Survey for use in chemistry. *J. Chem. Educ.*, 85(10):1435, 2008.

[207] K. Semsar, J.K. Knight, G. Birol, and M.K. Smith. The Colorado Learning Attitudes about Science Survey (CLASS) for use in biology. *CBE-Life Sci. Educ.*, 10(3):268–278, 2011.

[208] L. Ding. Verification of causal influences of reasoning skills and epistemology on physics conceptual learning. *Phys. Rev. ST Phys. Educ. Res.*, 10(2):023101.

[209] A. Traxler and E. Brewe. Equity investigation of attitudinal shifts in introductory physics. *Phys. Rev. ST Phys. Educ. Res.*, 11(2):020132, 2015.

[210] C. Baily and N.D. Finkelstein. Development of quantum perspectives in modern physics. *Phys. Rev. ST Phys. Educ. Res.*, 5:010106, Mar 2009.

[211] E. Gire, B. Jones, and E. Price. Characterizing the epistemological development of physics majors. *Phys. Rev. ST Phys. Educ. Res.*, 5(1):010103, 2009.

[212] B.M. Zwickl, N. Finkelstein, and H.J. Lewandowski. Development and validation of the Colorado Learning Attitudes about Science Survey for experimental physics. *AIP Conf. Proc.*, 1513(1):442–445, 2013.

[213] R. Wulf, L.M. Mayhew, and N.D. Finkelstein. Impact of informal science education on children's attitudes about science. *AIP Conf. Proc.*, 1289(1):337–340, 2010.

[214] V. Sawtelle, E. Brewe, and L. Kramer. Validation study of the Colorado Learning Attitudes about Science Survey at a Hispanic-serving institution. *Phys. Rev. ST Phys. Educ. Res.*, 5(2):023101, 2009.

[215] I. Kontro and D. Buschhüter. Validity of Colorado Learning Attitudes about Science Survey for a high-achieving, finnish population. *Phys. Rev. Phys. Educ. Res.*, 16(2):020104, 2020.

[216] M.W. Watkins. *A Step-by-step Guide to Exploratory Factor Analysis with SPSS*. Routledge, New York, NY, 2021.

[217] D. Çokluk B.O., Koçak. Using Horn's parallel analysis method in exploratory factor analysis for determining the number of factors. *Educ. Sci.-Theory Prac.*, 16, 2016.

[218] S. Sharma, S. Mukherjee, A. Kumar, and W.R. Dillon. A simulation study to investigate the use of cutoff values for assessing model fit in covariance structure models. *J. Busi. Res.*, 58(7):935–943, 2005.

[219] V. Irwin, J. Zhang, X. Wang, S. Hein, K. Wang, A. Roberts, C. York, A. Barmer, F. Bullock Mann, R. Dilig, et al. Report on the Condition of Education 2021. NCES 2021-144. *Nat. Cen. Educ. Stat.*, 2021.

[220] R. Scherr. Never mind the gap: Gender-related research in Physical Review Physics Education Research, 2005–2016. *Phys. Rev. Phys. Educ. Res.*, 12(2):020003, 2016.

[221] R. Henderson and J. Stewart. Racial and ethnic bias in the Force Concept Inventory. *Phys. Educ. Res. Conf. Proceedings*, 2017.

[222] R. Henderson, C. Zabriskie, and J. Stewart. Rural and first generation performance differences on the Force and Motion Conceptual Evaluation. In *Physics Education Research Conference 2017*, PER Conference, page 172, Washington, DC, July 26-27 2018.

[223] D.S. Hewagallage, J. Stewart, and R. Henderson. Differences in the predictive power of pretest scores of students underrepresented in physics. In *Phys. Educ. Res. Conf. 2019*, page 172, College Park, MD, 2019. AIP.

[224] J.M. Nissen. Gender differences in self-efficacy states in high school physics. *Phys. Rev. Phys. Educ. Res.*, 15(1):013102, 2019.

[225] 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.

[226] A.M.L. Cavallo, W.H. Potter, and M. Rozman. Gender differences in learning constructs, shifts in learning constructs, and their relationship to course achievement in a structured inquiry, yearlong college physics course for life science majors. *School Sci. Math.*, 104(6):288, 2004.

[227] C. Lindstrøm and M.D. Sharma. Self-efficacy of first year university physics students: Do gender and prior formal instruction in physics matter? *Int. J. Innov. Sci. Math. Educ.*, 19(2), 2011.

[228] L.E. Kost-Smith. *Characterizing, Modeling, and Addressing Gender Disparities in Introductory College Physics*. PhD thesis, University of Colorado Boulder, 2011.

[229] V. Sawtelle, E. Brewe, and L.H. Kramer. Exploring the relationship between self-efficacy and retention in introductory physics. *J. Res. Sci. Teach.*, 49(9):1096, 2012.

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

[231] E.M. Marshman, Z.Y. Kalender, T. Nokes-Malach, C. Schunn, and C. Singh. Female students with A's have similar physics self-efficacy as male students with C's in introductory courses: A cause for alarm? *Phys. Rev. Phys. Educ. Res.*, 14(2):020123, 2018.

[232] Z.Y. Kalender, E. Marshman, Christian D. Schunn, T.J. Nokes-Malach, and C. Singh. Damage caused by women's lower self-efficacy on physics learning. *Phys. Rev. Phys. Educ. Res.*, 16:010118, Apr 2020.

[233] C. Huang. Gender differences in academic self-efficacy: a meta-analysis. *Eur. J. Psychol. Educ.*, 28(1):1, 2013.

[234] J.M. Hall and M.K. Ponton. Mathematics self-efficacy of college freshman. *J. Dev. Educ.*, 28(3):26, 2005.

[235] M.L. Peters. Examining the relationships among classroom climate, self-efficacy, and achievement in undergraduate mathematics: A multi-level analysis. *Int. J. Sci. Math. Educ.*, 11(2):459, 2013.

[236] M.L. Bernacki, T.J. Nokes-Malach, and V. Aleven. Examining self-efficacy during learning: Variability and relations to behavior, performance, and learning. *Metacogn. Learn.*, 10:99, 2015.

[237] H. Hartman and M. Hartman. Do gender differences in undergraduate engineering orientations persist when major is controlled? *Int. J. Gender Sci. Tech.*, 1(1), 2009.

[238] M. Micari, P. Pazos, and M. J.Z. Hartmann. A matter of confidence: Gender differences in attitudes toward engaging in lab and course work in undergraduate engineering. *J. Women Minorities Sci. Eng.*, 13(3), 2007.

[239] E. Cech, B. Rubineau, S. Silbey, and C. Seron. Professional role confidence and gendered persistence in engineering. *Am. Sociol. Rev.*, 76(5):641, 2011.

[240] J.P. Concannon and L.H. Barrow. A reanalysis of engineering majors' self-efficacy beliefs. *J. Sci. Educ. Technol.*, 21(6):742, 2012.

[241] C.M. Jagacinski. Women engineering students: Competence perceptions and achievement goals in the freshman engineering course. *Sex Roles*, 69(11-12):644, 2013.

[242] R.W. Lent, S.D. Brown, H. Sheu, J. Schmidt, B.R. Brenner, C.S. Gloster, G. Wilkins, L.C. Schmidt, H. Lyons, and D. Treistman. Social cognitive predictors of academic interests and goals in engineering: Utility for women and students at historically Black universities. *J. Couns. Psychol.*, 52(1):84, 2005.

[243] M.A. Hutchison, D.K. Follman, M.B. Sumpter, and M. George. Factors influencing the self-efficacy beliefs of first-year engineering students. *J. Eng. Educ.*, 95(1):39, 2006.

[244] R.W. Lent, H.B. Sheu, D. Singley, J.A. Schmidt, L.C. Schmidt, and C.S. Gloster. Longitudinal relations of self-efficacy to outcome expectations, interests, and major choice goals in engineering students. *J. Vocat. Behav.*, 73(2):328, 2008.

[245] A. Uitto. Interests, attitudes and self-efficacy beliefs explaining upper-secondary school students' orientation towards biology-related careers. *Int. J. Sci. Math. Educ.*, 12(6):1425, 2014.

[246] L. Ainscough, E. Foulis, K. Colthorpe, K. Zimbardi, M. Robertson-Dean, P. Chunduri, and L. Lluka. Changes in biology self-efficacy during a first-year university course. *CBE—Life Sci. Educ.*, 15(2):1, 2016.

[247] J. Dalgety and R.K. Coll. Exploring first-year science students' chemistry self-efficacy. *Int. J. Sci. Math. Educ.*, 4(1):97, 2006.

[248] S.M. Villafañe, C.A. Garcia, and J.E. Lewis. Exploring diverse students' trends in chemistry self-efficacy throughout a semester of college-level preparatory chemistry. *Chem. Educ. Res. Pract.*, 15(2):114, 2014.

[249] G.V. Caprara, M. Vecchione, G. Alessandri, M. Gerbino, and C. Barbaranelli. The contribution of personality traits and self-efficacy beliefs to academic achievement: A longitudinal study. *Brit. J. Educ. Psychol.*, 81(1):78, 2011.

[250] K.D. Multon, S.D. Brown, and Robert W. Lent. Relation of self-efficacy beliefs to academic outcomes: A meta-analytic investigation. *J. Couns. Psychol.*, 38(1):30, 1991.

[251] T.A. Judge and R. Ilies. Relationship of personality to performance motivation: A meta-analytic review. *J. Appl. Psychol.*, 87(4):797, 2002.

[252] I. Sanchez-Cardona, R. Rodriguez-Montalbán, E. Acevedo-Soto, K.N. Lugo, F. Torres-Oquendo, and J. Toro-Alfonso. Self-efficacy and openness to experience as antecedent of study engagement: An exploratory analysis. *Procedia Soc. Behav. Sci.*, 46:2163, 2012.

[253] N. Schmitt. The interaction of neuroticism and gender and its impact on self-efficacy and performance. *Hum. Perform.*, 21(1):49, 2007.

[254] G. Chen, W.J. Casper, and J.M. Cortina. The roles of self-efficacy and task complexity in the relationships among cognitive ability, conscientiousness, and work-related performance: A meta-analytic examination. *Hum. Perform.*, 14(3):209, 2001.

[255] L. Dörrenbächer and F. Perels. Self-regulated learning profiles in college students: Their relationship to achievement, personality, and the effectiveness of an intervention to foster self-regulated learning. *Learn. Individ. Differ.*, 51:229, 2016.

[256] C.C. Cohen and N. Deterding. Widening the net: National estimates of gender disparities in engineering. *J. Eng. Educ.*, 98(3):211, 2009.

[257] J.M. Braxton, W.R. Doyle, H.V. Hartley III, A.S. Hirschy, W.A. Jones, and M.K. McLendon. *Rethinking College Student Retention*. John Wiley & Sons, San Francisco, CA, 2013.

[258] R.A. Alvaro. *The Effectiveness of a Science Therapy Program upon Science Anxious Undergraduates*. PhD thesis, Loyola University Chicago, 1978.

[259] J.V. Mallow. A science anxiety program. *Am. J. Phys.*, 46(8):862–862, 1978.

[260] K. Williams. Understanding communication anxiety and gender in physics. *J. Coll. Sci. Teach.*, 30(4):232, 2000.

[261] N. Hall and D. Webb. Instructor's support of student autonomy in an introductory physics course. *Phys. Rev. Phys. Educ. Res.*, 10(2):020116, 2014.

[262] R. W. Lent, S. D. Brown, and G. Hackett. Toward a unifying social cognitive theory of career and academic interest, choice, and performance. *J. Vocat. Behav.*, 45(1):79–122, 1994.

[263] A. Diseth. Self-efficacy, goal orientations and learning strategies as mediators between preceding and subsequent academic achievement. *Learn. Individ. Differ.*, 21(2):191, 2011.

[264] W.K. Hofstee, B. de Raad, and L.R. Goldberg. Integration of the Big Five and circumplex approaches to trait structure. *J. Abnorm. Psychol. Soc. Psychol.*, 63(1):146, 1992.

[265] J. Stewart, R. Henderson, L. Michaluk, J. Deshler, E. Fuller, and K. Rambo-Hernandez. Using the social cognitive theory framework to chart gender differences in the developmental trajectory of STEM self-efficacy in science and engineering students. *J. Sci. Educ. Tech.*, 29(6):758–773, 2020.

[266] A.L. Traxler, X.C. Cid, J. Blue, and R. Barthelemy. Enriching gender in physics education research: A binary past and a complex future. *Phys. Rev. Phys. Educ. Res.*, 12:020114, Aug 2016.

[267] S. Srivastava, O.P. John, S.D. Gosling, and J. Potter. Development of personality in early and middle adulthood: Set like plaster or persistent change? *J. Pers. Soc. Psychol.*, 84(5):1041, 2003.