Investigating Gender Differences in Test Anxiety, Self-efficacy, Mindset, Grade Penalty and Grades in Physics Courses: A Quest for Equity

by

Alysa Malespina

B.A., Rollins College, 2018

M.S., University of Pittsburgh, 2023

Submitted to the Graduate Faculty of

the Dietrich School of Arts and Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2023

UNIVERSITY OF PITTSBURGH

DIETRICH SCHOOL OF ARTS AND SCIENCES

This dissertation was presented

by

Alysa Malespina

It was defended on

July 27th 2023

and approved by

Dr. Russell Clark, Senior Lecturer, Department of Physics Astronomy

Dr. Robert Devaty, Associate Professor, Department of Physics and Astronomy

Dr. Roger Mong, Associate Professor, Department of Physics and Astronomy

Dr. Christian Schunn, Professor, Department of Psychology

Thesis Advisor/Dissertation Director: Dr. Chandralekha Singh, Distinguished Professor, Department of Physics and Astronomy

Investigating Gender Differences in Test Anxiety, Self-efficacy, Mindset, Grade Penalty and Grades in Physics Courses: A Quest for Equity

Alysa Malespina, PhD

University of Pittsburgh, 2023

Students' grades and motivational beliefs about physics can influence their performance and persistence in science, technology, engineering, and math (STEM) disciplines, as well as their future career opportunities and goals. In recent years, many studies have used these outcomes as measures of equity in physics classrooms. Students from traditionally marginalized groups in physics (such as women) may not have the support and resources needed to develop strong motivational beliefs in physics. They have to contend with societal stereotypes and biases about who can excel in physics throughout their lives and are less likely to take advanced physics in high school. In this dissertation, I investigate the relationship between gender, physics motivational beliefs and grade outcomes for students.

Through my quantitative studies, I first analyzed gender differences in students' physics self-efficacy and test anxiety and how those constructs predict high-stakes and low-stakes test performance. Next, I investigated how perception of the effectiveness of peer interactions can influence women and men's physics self-efficacy, and how these measures predict performance. Additionally, I investigated gender differences in physics intelligence mindset and analyzed how mindset predicted course grades. Lastly, I investigated gender differences in grade penalties (grade penalty for a group is defined as a lower grade in a course compared to the overall grade point average up to that point).

These findings can be useful to instructors who aim to make their courses more equitable and inclusive for all students. I discuss approaches that can make the learning environment more equitable and inclusive while maintaining high standards.

Table of Contents

List of Tables

List of Figures

Preface

First, I would like to thank my research advisor, Dr. Chandralekha Singh for her support throughout my studies. Her passion for equity in physics education and encouragement to explore new constructs shaped my research for the better.

I am grateful to Dr. Russell Clark and Dr. Danny Doucette for helping me start my journey as an instructor and giving me the confidence to share my love of teaching with others. I would like to thank Dr. Christian Schunn for helping me understand how to use and interpret the statistical methods I use in my research. I thank Dr. Robert Devaty for his generous help making every manuscript I wrote better. I would like to thank Dr. Roger Mong for his insightful feedback.

Next, I would like to thank my professors at Rollins College, especially Dr. Anne Murdaugh in the Physics Department and Dr. Ryan Musgrave in the Philosophy Department. They encouraged all of my intellectual interests (no matter how relevant to their classes), which led me to physics education research.

I would also like to thank my parents, as well as the friends I've made in Pittsburgh over the past five years: Bob and Lauren Caddy, Mary Jane Brundage, Lisabeth Santana, and Si Wang. I especially thank Doris Li and Sonja Cwik who I shared a research group and graduate cohort with, for their friendship, support, and company over the years. I thank my partner, Arcelia Posadas for her unwavering support and for never doubting I could succeed.

Lastly, while I don't have room to thank everyone who made this dissertation possible, I would like to thank the students, friends, and colleagues that helped me along the way.

1.0 Introduction

Women are underrepresented in many science, technology, engineering, and mathematics (STEM) disciplines [1–4]. These gender disparities are particularly large in fields such as physics, in which only one-fifth of degree-earning undergraduates are women [1, 5]. Additionally, physics courses often have gender differences in course grades and motivational factors.

Gendered grade differences are important to address in part because grades affect students' access for scholarships, graduate and professional school admissions, and career opportunities. Gender differences in motivational factors are also important for many reasons. For example, they are related to performance and persistence in physics [6–12]. Prior research from our research group has investigated the gender differences in physics performance and motivational beliefs across a variety of contexts [**?**, 3, 9, 10, 13–50].

Representation and retention of women in physics, who have historically been excluded, is important to ensure that we do not miss out on the talents of this pool of individuals even in the twenty first century.

If physics departments do not adequately support and retain women by creating an equitable and inclusive environment, they will lose the talent of many women who may have excelled in the major and future physics careers but decided against doing so due to the current chilly climate and culture of physics. In prior studies, gendered grade and motivational belief differences have been observed in engineering and other physical science majors [51,52]. Women leave physics the major at a higher rate than men, and often cite concerns over low grades as a reason for changing their major [53]. Some studies have found that women who leave physics and other STEM majors tend to have higher grades than the men who leave, and sometimes have grades comparable to the men who remain in the major [39].

The importance of equitable outcomes based upon students' gender holds also for students who are not physics majors, such as engineering and bioscience majors. While the students in these programs are unlikely to become physicists because they usually have aspirations to pursue engineering or health-related careers, introductory physics courses are mandatory for these groups at least in the US. Large-enrollment introductory science courses often act as "weed-out" courses for students, and may discourage students from pursuing other STEM disciplines [53,54]. In physics courses for bioscience majors, women are not underrepresented, whereas in physics courses for engineering and physical science students women are underrepresented. One important reason for investigating physics courses for both engineering/physical science majors and bioscience majors is to compare gender differences in these two contexts.

Gender differences in grades and motivational factors have often been attributed to unequal opportunities to participate in physics as well as pervasive societal stereotypes and biases about who can excel in physics [55–58]. For example, fields that are widely perceived to require innate talent or "brillinace", such as physics, tend to have less gender diversity than fields that do not have associations with brilliance [59]. Efforts in increase women's opportunities to flourish and persist in physics are important both to the women taking physics courses, and the departments that have the opportunity to recruit and retain these women's interests and talents.

1.1 Motivational Beliefs

Part of this research presented here focuses on test anxiety, self-efficacy, intelligence mindset, and academic self-concept.

Test anxiety is a phenomenon with several facets. The cognitive facet consists of worry and self-preoccupation (e.g., thinking about one's perceived shortcomings instead of the task at hand), as well as intrusive thoughts of failure, all of which limit the time and cognitive resources students with test anxiety can devote to the assessment [60, 61]. The affective facet of test anxiety affects how students feel when they have test anxiety, for example, a fast heartbeat or "butterflies in their stomach" [60]. The behavioral aspect manifests in avoidance techniques, such as procrastination or interacting only with surface-level feedback after the exam (e.g., not examining mistakes closely to make a plan for future improvement) [60, 62]. It is possible for students to experience all three facets, or only a subset of them [60]. In addition, women are more likely to report test anxiety than men [60, 63], so understanding test anxiety is vital to create equitable learning environments.

Self-efficacy is one's belief in their capability to succeed at an activity or subject [63, 64], and has been linked to positive outcomes for physics students [11, 12, 40, 41, 65]. Self-efficacy is developed through four mechanisms: mastery experiences, such as overcoming obstacles; social modeling, or people similar to oneself succeeding; social persuasion, encouragement to increase resolve and measure success via personal improvement; and emotional states, such as management of anxiety [64]. Female students tend to have lower self-efficacy than male students in physics [3, 11, 12, 40, 41, 65]. Prior research has attributed this to many related factors alluded to earlier, e.g., women are less likely to have the same level of previous mastery experiences as men (because they are less likely to take advanced physics in high school [66]), they have few role models due to under-representation of women in physics [3, 66], and they are less likely to receive encouragement that they can succeed in physics from family [67], instructors [3, 28], and society at large [55]. In addition to investigating self-efficacy directly, we also introduce a new measure: self-reported "peer influence on self-efficacy", which measures students' perceptions about how interactions with their peers affected their confidence in physics. This measure may be especially useful for high-enrollment classes in which students interact much more with other students than with the instructor or teaching assistant.

Intelligence mindset describes a person's beliefs about the nature of intelligence [68]. A growth mindset is one in which intelligence is viewed as something that can be cultivated with effort, like a muscle, while a fixed mindset is one in which intelligence is thought to be innate and unchangeable [68]. The mindsets held by learners are thought to shape how students engage in learning. With a fixed mindset, a student is likely to disengage from or avoid difficult tasks [68–71]. On the other hand, the engagement, propensity to attempt challenging problems, and persistence that often come with growth mindsets have been linked to positive learning outcomes [68–71], even after controlling for prior academic achievement [68, 72–75]. Growth mindsets have also been linked to greater participation in STEM fields for students from underrepresented groups [76], and can be a useful resource for underrepresented students to combat stereotype threat or anxiety [77].

Expectancy Value Theory is a framework to understand student achievement and persistence in a domain [78–80]. Expectancy Value Theory posits that performance and persistence is determined by someone's expectation of success and the extent to which they value that task. Expectation of success relies on factors such as academic self-concept. Academic self-concept describes a long-term expectation of success that students hold regarding their academic abilities and that primarily depends on grades and outside feedback (e.g., from parents, peers, and instructors) [78]. Grades inform academic self-concept as both an external ("How good at math am I compared to other students?") and internal ("How good am I at math compared to English?") frame of reference [78]. Low academic self-concept may lead to lower future achievement and persistence because it discourages student engagement in a domain [79].

1.2 Overview

In chapter 2 we focus on female and male students' self-efficacy and test anxiety in introductory physics courses for engineering and physical science majors. We explore the relationships between self-efficacy, test anxiety, and gender differences in introductory calculus-based physics performance. Although there has been research that uses test anxiety and self-efficacy to predict student grades, no study to our knowledge has investigated this in the context of low- (e.g., homework and quizzes) and high-stakes (e.g., traditional exams) physics assessments. Using validated survey data and grade information, we compared the predictive power of self-efficacy and test anxiety on student performance on a variety of assessment types. We found that there are gender differences in both self-efficacy and test anxiety, as well as in high-stakes assessment outcomes. There were no gender differences in low-stakes assessment scores. Further, we found that models that control for self-efficacy and/or test anxiety eliminate the predictive power of gender for high-stakes assessment outcomes. Finally, we found that self-efficacy partially mediates the effect of test anxiety on high-stakes assessment outcomes. From these results, we make several suggestions for instructors that may alleviate the adverse effects of test anxiety and make physics assessments more equitable and inclusive.

In chapter 3, we investigate the relationship between self-efficacy, test anxiety, and gender differences in performance in an introductory physics sequence for bioscience students. Using validated survey data and grade information from students in a two-semester introductory physics course sequence, we compared the predictive power of self-efficacy and test anxiety on student performance on both low- and high-stakes assessments. We found that there are gender differences in self-efficacy, test anxiety, and high-stakes assessment outcomes in both Physics 1 and Physics 2. There were no gender differences in low-stakes assessment scores. We also found that self-efficacy and test anxiety predicted high-stakes (but not low-stakes) assessment outcomes in both Physics 1 and Physics 2.

In chapter 4, we explore differences in motivational factors and learning outcomes between students in introductory physics courses who took online classes during remote instruction due to COVID-19 and those who took in-person classes. We first investigated mean differences in students' self-efficacy, test anxiety, and learning outcomes in two categories: low-stakes (homework, quizzes) and high-stakes (exams) assessments. We found that most differences were small or moderate; however, students performed drastically better on exams during remote classes compared to in-person classes. This may be partially attributed to different exam formats for remote versus in-person classes. Gender differences in high-stakes assessment grades were also eliminated during online instruction. Finally, we find that in both in-person and remote courses, test anxiety predicts self-efficacy, which in turn predicts high-stakes assessment outcomes. From these results, we make several suggestions for instructors that may alleviate the adverse effects of test anxiety and make physics assessments more equitable and inclusive.

In chapter 5, we describe a study in which a validated motivational survey was used to investigate the effect of working in mixed or same-gender groups on physics self-efficacy and self-reported "peer influence on self-efficacy" in a calculus-based introductory physics course in which women are severely underrepresented both in our sample and in the US broadly. Partly due to societal stereotypes and biases, we found that men tended to have higher physics self-efficacy and reported higher peer influence on self-efficacy than women both before and at the end of the physics course. Additionally, all students except those who worked in same-gender groups had a decrease in average physics self-efficacy from the beginning to the end of the semester. Finally, using mediation analysis, we found that gender predicted self-efficacy for students who worked in mixed gender groups, but not for those in same-gender groups. Our findings suggest that instructors should implement classroom policies that encourage equitable and inclusive group work, so that all students can thrive.

In chapter 6, we investigate intelligence mindset (i.e., the belief that intelligence is either innate and unchangeable or can be developed). We studied 781 students in calculus-based Physics 1 to investigate if their mindset views were separable into more nuanced dimensions, if they varied by gender/sex and over time, and if they predicted course grade. Confirmatory factor analysis was used to divide mindset survey questions along two dimensions: myself versus others and growth versus ability aspects of mindset. Paired and unpaired t-tests were used to compare mindset factors over time and between genders, respectively. Multiple regression analysis was used to find which mindset factors were the best predictors of course grade. This study shows that intelligence mindset can be divided into four factors: My Ability, My Growth, Others' Ability, and Others' Growth. Further, it reveals that gender differences are more pronounced in the "My" categories than the "Others'" categories. At the start of the course, there are no gender differences in any mindset component, except for My Ability. However, gender differences develop in each component from the start to the end of the course, and in the My Ability category, the gender differences increase over time. Finally, we find that My Ability is the only mindset factor that predicts course grade. These results allow for a more nuanced view of intelligence mindset than has been suggested in previous interview and survey-based work. By investigating the differences in mindset factors over time, we see that learning environments affect women's and men's intelligence mindsets differently. The largest gender difference is in My Ability, the factor that best predicts course grade. This finding has implications for developing future mindset interventions and opens new opportunities to eliminate classroom inequities.

In chapter 7, we study how bioscience students' motivational beliefs, such as disciplinary intelligence mindsets, can influence their physics performance and persistence. Intelligence mindset beliefs have long been argued to fall along a continuum between fixed and growth mindsets. Those with fixed physics mindsets believe that ability in physics is innate and unchangeable, while those with growth mindset believe that ability in physics can be developed with effort. More recent research with physical science and engineering majors suggests that these are somewhat separable beliefs, with some students believing aspects of both fixed and growth mindsets, and that students can hold different beliefs about other students vs. beliefs about themselves (e.g., others could improve through effort but they themselves could not). In this study, 419 students in Physics 1 for students pursuing bioscience majors took pre- and post- physics mindset surveys. We investigated whether the physics mindset views of students pursuing bioscience or health-related majors were separable into more nuanced dimensions, if the means and distribution of these views varied by gender/sex and over time, and if any of these views predicted course grade. Replicating prior findings with physical science and engineering majors, we found that intelligence mindsets can be divided into four separable but correlated constructs: My Ability, My Growth, Others' Ability, and Others' Growth. Further, in this bioscience/health-related majors group, the "Ability" beliefs grew stronger and the "Growth" beliefs became weaker over time. These shifts were particularly strong for women. The changes in beliefs were also stronger for "My" beliefs than "Others'" beliefs for both men and women. My Ability and My Growth scores were also the strongest predictors of course grades above and beyond academic preparation differences as assessed by high school GPA and SAT/ACT Math scores. These findings have implications for eliminating classroom inequities.

In chapter 8, we introduce a framework that posits that grade penalty is a measure of academic self-concept and investigate if there are gender differences in grade penalties in physics courses for students majoring in physics. In order to quantify grade penalty, we define grade anomaly as the difference between a student's grade in a course under consideration and their grade point average (GPA) in all other classes thus far. A grade anomaly lower than students' expected grade based on their GPA is a grade penalty and higher than expected average grade is a grade bonus. Our framework posits that since women have traditionally been marginalized in physics, female physics majors are more likely to be negatively impacted by a grade penalty in their courses since their academic self-concept as a physics major hinges on them securing a certain grade. In the study presented here, we examine the average grade anomalies across a number of courses for female and male physics majors. We find that these students received grade penalties in almost all physics courses studied, though there were grade bonuses in a few laboratory courses. We also find that in physics courses, on average, women often had larger grade penalties than men, especially in introductory courses. We hypothesize that, because their grade penalties are often larger than men's, women's decisions to pursue a physics major and career may be particularly affected by grade penalties received in their various courses. Furthermore, the grade penalty measure can be easily computed by the physics programs concerned with equity.

In chapter 9, we continue to use the average grade anomaly framework that posits that grade penalty in first year foundational science courses for engineering majors may be particularly damaging to female students who do not have role models and are questioning whether they have what it takes to excel in an engineering major and career due to pervasive stereotypes. In order to quantify grade penalty, we define Grade Anomaly as the difference between a student's grade in a course under consideration and their grade point average (GPA) in all other classes thus far. A grade anomaly lower than students' expected grade based on their GPA is a grade penalty and higher than expected grade is a grade bonus. Our framework posits that female engineering majors are more likely to be negatively impacted by a grade penalty in their first-year foundational science courses since their academic self-concept as an engineering major hinges on them securing a certain grade. In the study presented here, we examine Average Grade Anomalies of 6,028 first-year engineering students across a number of required courses. We find that students tend to receive grade bonuses in engineering and English composition courses, and grade penalties in physics, chemistry, and math courses. These courses with grade penalties tend to be large, lecture-based courses. We also find that in physics courses, women have larger grade penalties than men, whereas in chemistry and math, men have larger grade penalties. Thus, physics courses may be most damaging to women out of all of the courses in which they receive grade penalty. We hypothesize that women's decisions to pursue an engineering major and career may be affected more by the grade penalty received in foundational science courses than men's due to societal stereotypes about who can excel in engineering and access to other coping mechanisms that may help to rationalize lower-than-expected grades. Furthermore, the grade penalty measure can be easily computed by the engineering programs concerned with equity. Finally, we provide recommendations for how engineering programs may mitigate grade penalties in the foundational science courses, which may be particularly damaging to women.

In chapter 10, we investigate a framework that posits that since women studying bioscience have traditionally been marginalized in the sciences, they are more likely to be negatively impacted by a grade penalty in their courses since their academic self-concept may be dependent on receiving a certain grade. In this study, we examined AGAs of 2,445 students across a number courses. We found that on average students received grade penalties in the twelve most commonly taken science courses for bioscience students at our institution. We also found that women had more extreme grade penalties than men in seven of the twelve science classes we investigated. We hypothesize that women's decisions to pursue STEM careers may be affected more by the grade penalty received in required science courses than men's because their grade penalties are often larger.

In chapter 11, we use grades and "grade anomalies" to investigate student performance before, during, and after the period of COVID-19 remote instruction in courses for first-year engineering majors. We also use these measures to investigate gender equity in these courses. We investigated all required courses for this group of students and found that the Engineering and English Composition courses tended to have grade bonuses, while Mathematics, Physics, and Chemistry courses tended to have grade penalties. We broadly find that both grades and grade penalties showed positive trends during remote instruction and deteriorated after remote instruction. We also find that there were many more gender differences in grade anomalies than in grades. We hypothesize that women's decisions to pursue STEM careers may be affected more by the grade penalty received in required science courses than men's because their grade penalties are often larger during all time periods studied.

In chapter 12, we use grades and "grade anomalies" to investigate student performance before, during, and after COVID-19 remote instruction in courses for bioscience and health-related majors. We also use these measures to investigate gender equity in these courses. Students received grade penalties in all courses studied, consisting of the twelve courses taken by the greatest number of bioscience and health-related majors. We broadly find that both grades and grade penalties showed positive trends during remote courses and deteriorated after remote instruction. We also find that there were many more gender differences in grade anomalies than in grades. We hypothesize that women's decisions to pursue STEM careers may be affected more by the grade penalty received in required science courses than men's because their grade penalties are often larger during all time periods studied.

Lastly, in the final chapter, we briefly discuss some future directions based upon this work.

2.0 Gender differences in test anxiety and self-efficacy: Why instructors should emphasize low-stakes formative assessments in physics courses

2.1 Introduction

Research on gender differences in introductory physics course performance is abundant, as is the work focusing on ways to mitigate this "gender gap" [81–84]. Some of this work focuses on how student experiences before they enter the classroom affect physics performance, and how societal changes could increase opportunities for women in science, technology, mathematics, and engineering (STEM) fields. Examples include societal stereotypes about who belongs in physics [23,55,67,85,86], and limited opportunities to take advanced physics courses in high school resulting in gendered differences in prior preparation [66]. Other work focuses on changes in the classroom that could make physics more equitable. Examples of this approach include addressing gender bias in exams and standardized tests [87–89] and investigating the effects of alternatives to lecture-based courses (such as evidence-based active learning) on gender differences in performance [90–92]. In this work, we use a third approach that studies the crossroads of in-class and out-of-class experiences: the study of motivational beliefs [9, 28] and how those beliefs affect classroom experiences. Previous research has shown that gendered performance differences can be attributed to differences in motivational beliefs about physics between male and female students (which can again at least partly be attributed to societal stereotypes about physics) [11, 12, 23, 32, 40, 41, 65, 93, 94].

Here, we study two factors that have been linked to gender differences in physics performance. The first is test anxiety (TA) [93, 94], a phenomenon that affects students' test performance and is more likely to affect women [63, 95]. The second is self-efficacy (SE) [11, 12, 40, 41, 65], a student's belief in their ability to complete a task [63, 64]. Because both TA and SE have been studied independently in the physics context, we are particularly interested in the interactions between the two, and if the type of assessment (e.g., homework or exams) is important when observing their effects. In this study, we investigate the relationship between gender, low- and high-stakes assessment outcomes, test anxiety, and self-efficacy. Specifically, we aim to answer the following questions:

RQ 1. Are there gender differences in students' prior preparation, self-efficacy, or test anxiety?

RQ 2. Are there gender differences in students' low- and high-stakes assessment scores?

RQ 3. Are self-efficacy and test anxiety independent predictors of assessment scores?

RQ 4. Does gender predict scores in low- or high-stakes assessments when controlling for self-efficacy or test anxiety?

2.1.1 Gender differences in Physics Courses

Performance differences between male and female students in physics courses are often due to sociocultural stereotypes and biases pertaining to who belongs in physics and who can excel in it, and insufficient efforts to counter them in order to make the learning environment more equitable and inclusive. For example, girls are less likely than boys to have parents who believe they can excel in the sciences so parents are less likely to encourage them to pursue related courses and activities from early on [96, 97]. This, combined with societal stereotypes that success in physics requires particular brilliance and brilliance is associated with men, in part explains the low numbers of women in the field [59]. Women are less likely than men to take physics in high school [66], so they are less likely to have prior experience if they are required to take physics in college. Once women are enrolled in physics courses, they tend to have lower self-efficacy, which is an important predictor of physics performance, even when controlling for prior academic preparation [11, 12, 40, 41, 65].

Though our focus here is on the relationship between text anxiety and self-efficacy in the physics context, prior work focusing on biology classes establishes a relationship between test anxiety and high-stakes assessment outcomes [94]. However, we chose to incorporate self-efficacy into the study for two reasons: first because the relationship between performance and self-efficacy is well-documented in physics [11,12,40,41,65], and because management of anxiety is explicitly mentioned as a mechanism to build self-efficacy [63].

2.1.2 Self-Efficacy

Self-efficacy is one's belief in their capability to succeed at an activity or subject [63, 64], and has been linked to positive outcomes for physics students [11, 12, 40, 41, 65]. Self-efficacy is developed through four mechanisms. The first is mastery experiences, which describes learning by overcoming difficulties such as a challenging homework assignment. The second is social modeling, or having role models. This describes seeing people similar (for example, somebody of a similar age, ethnicity, or gender) to oneself succeeding in a domain. The third is social persuasion, which is encouragement to increase resolve and measure success via personal improvement. Though it is not necessary for a potential role model to share all of a student's identities, prior work has shown that women especially benefit from other women as peers and role models [98, 99]. The final mechanism is emotional state, such as management of anxiety [64]. Because they have fewer opportunities to utilize the first three of these mechanisms, female students tend to have lower self-efficacy than male students in physics [11, 12, 40, 41, 65]. For example, women are less likely to have the same level of previous mastery experiences as men (because they are less likely to take advanced physics courses in high school for a variety of reasons [66]), they have few role models due to under-representation of women in physics [66], and they are less likely to receive encouragement that they can succeed in physics from family [67], instructors [28], and society at large [55]. Because self-efficacy allows students to develop coping mechanisms that could thwart test anxiety [63], we hypothesize that students with high self-efficacy will also have low test-anxiety.

To explore how test anxiety and self-efficacy predict students' performance in different situations, we compare female and male students' performance on low-stakes and high-stakes assessments. Here, low-stakes assessments are those that make up a small portion of a student's grade, such as recitation quizzes. High-stakes assessments are individual assessments that make up a large portion of a student's grade (e.g., traditional exams [100]). In this paradigm, ten assessments that each make up five percent of a student's grade would be relatively low-stakes, while a single assessment that makes up fifty percent of a student's grade would be high-stakes, even if the content was identical between the assessments. Because previous studies have shown larger gender gaps in high-stakes than low-stakes assessment [100], and women are more likely to report low self-efficacy [41] and high test anxiety [63, 95] than men, we hypothesize that test anxiety and self-efficacy will more strongly predict high-stakes than low-stakes assessment outcomes.

2.1.3 Test Anxiety

Test anxiety is a phenomenon with three facets: cognitive, affective, and behavioral. The cognitive facet consists of worry and self-preoccupation (e.g., thinking about one's perceived shortcomings instead of the task at hand), as well as intrusive thoughts of failure, all of which limit the time and cognitive resources students with test anxiety can devote to the assessment [95]. The affective facet of TA affects how students feel when they have test anxiety, for example, a fast heartbeat or "butterflies in their stomach" [95]. The behavioral aspect of test anxiety manifests in avoidance techniques, such as procrastination or interacting only with surface-level feedback after the exam (e.g., not examining mistakes closely to make a plan for future improvement) [95, 101].

Test anxiety can affect students in different ways: one student may have strong study skills but be unable to concentrate due to physiological manifestations of test anxiety, another may feel extremely anxious because they lack study skills to succeed in exams, while other students will have their own unique manifestations of test anxiety [95]. Ideally, educators will find ways to provide all students the environment they need to succeed in assessments without the burden of test anxiety, but this may require a multifaceted approach to meet different students' needs. Women are more likely to report test anxiety than men [63, 95], so understanding test anxiety and how to minimize its effect on student success is vital to create equitable learning environments.

Previous work in physics [93] has found that women report higher levels of TA than men, and that the predictive power of SE on the Force Concept Inventory (FCI) superseded that of TA. We aim to build upon these findings in two ways. First, instead of using the FCI and course exams, we compare course exams and "low-stakes" assessments, like homework and quizzes, so that results are more likely to generalize to courses that do not utilize the FCI. Second, we explore the interactions of TA and SE. We predict that, because the management of anxiety can contribute to high self-efficacy (and vice versa) [63], test-anxiety may predict self-efficacy.

2.2 Methodology

2.2.1 Participants and Procedures

This study took place at a large research university in the United States. We administered a multiple-choice motivational survey to students in an introductory calculus-based Physics 1 course. The surveys were given during the first and last week of their mandatory teaching assistant-led recitations. We call the first and final data sets "pre" and "post", respectively. For analysis, we longitudinally matched students ($N = 176$), including only students who completed both surveys and successfully passed an attention check (a question that requested the students select "C") on both. This research was carried out in accordance with the principles outlined in University of Pittsburgh Institutional Review Board (IRB) ethical policy. The traditional lecture-based course was taught by four instructors and primarily covered Newtonian mechanics. Students were either given extra credit or a participation grade for taking the survey, depending on the instructor.

Demographic data indicated our sample was 37% women. Students identified with the following races/ethnicities: 72% White, 14% Asian, 7% Hispanic/Latinx, 4% multiracial, 2% African American/Black, and 1% unspecified. The majority (80%) of students in the sample were first-semester first year students, and 60% of them were engineering majors. Outside of engineering, most students were undeclared or studying a physical science.

2.2.2 Measures

2.2.2.1 Prior Academic performance

We used high school grade point averages (HS GPAs) and Scholastic Achievement Test (SAT) scores to measure prior academic preparation. HS GPAs were reported on a 4.0 scale, and students with HS GPAs greater than 5.0 ($\sim 1\%$ of the sample) were excluded from analysis because their schools likely used a different grading system. SAT scores are on a scale of 200–800. American College Testing (ACT) scores were converted into approximate SAT scores [102]. If the student took a test more than once, the university provided the highest subject scores for the SAT or the highest composite score for the ACT. Both demographic data and prior academic information were acquired from de-identified university records via an honest broker.

2.2.2.2 Physics 1 Grades

Academic performance measures were provided by instructors. These measures included assessment scores for homework, quizzes, midterm exams, and final exams. For analysis, we divided these into two groups: "low-" and "high-stakes" assessments. The high-stakes assessment measure combined midterm and final exam grades. Because each instructor had three midterms and one final, each exam made up 25% of the high-stakes category score. The low-stakes assessment measure combined quiz and homework grades, and each assessment type counted for 50% of the category. Homework grades were only available for 74% of students. Students without homework grades were excluded from "low-stakes" analysis, though all of our findings were similar when they were included in the same analysis using only quiz scores. All other assessment scores were available for the whole sample. Before averaging, all assignments were normalized to a 10-point scale.

2.2.2.3 Survey

The TA survey questions were adapted from the previously validated [103, 104] Motivated Strategies for Learning Questionnaire [105]. To ensure we were measuring domain-specific mindset, we explicitly mentioned physics in the survey items, as seen in Table 1. For example "I have an uneasy, upset feeling when I take an exam," becomes "I have an uneasy, upset feeling when I take a physics test". SE survey questions were constructed from other surveys and previously validated [40, 41]. We further validated the survey through twenty one-hour student interviews to ensure that students interpreted questions as intended. Additionally, we performed confirmatory factor analysis (CFA) using the students in this study as a check for continued validity. For both the pre and post-surveys, the Comparative Fit Index (CFI) and Tucker Lewis Index (TLI) were ≥ 0.95 [106], the Root Mean Square Error of Approximation (RMSEA) was ≤ 0.08 [107], and the Standardized Root Mean Square Residual (SRMR) was ≤ 0.06 [106]. Standardized factor loadings ranged from 0.62-0.88 [106]. Chronbach's α was 0.75 and 0.81 for self-efficacy pre and post, while α was 0.88 and 0.90 for test anxiety pre and post. We found Pearson correlations between all variables. The weakest correlations were non-significant, while the strongest correlation was $r = 0.60$ ($p < 0.001$), between post test anxiety and post self-efficacy.

Test anxiety items were on a five-point Likert scale scale (1—Not at all true, 2—A

Table 1: The physics test anxiety (TA) and self-efficacy (SE) items on the survey. One item mentions a laboratory because this survey is used in multiple course contexts, not because the course in question has a laboratory component.

Item No.	Item Text
TA 1	I am so nervous during a physics test that I cannot remember what I have learned.
TA 2	I have an uneasy, upset feeling when I take a physics test.
TA 3	I worry a great deal about physics tests.
TA 4	When I take a physics test, I think about how poorly I am doing.
SE 1	I am able to help my classmates with physics in the laboratory or in recitation.
SE 2	I understand concepts I have studied in physics.
SE 3	If I study, I will do well on a physics test.
SE 4	If I encounter a setback in a physics exam, I can overcome it.

little true, 3—Somewhat true, 4—Mostly true, 5—Completely true) and self-efficacy questions were on a four-point Likert scale (1—NO!, 2—no, 3—yes, 4—YES!). A higher score that indicates the student has more test anxiety or higher self-efficacy. Thus, an ideal course outcome is that all students have low test anxiety scores and high self-efficacy scores.

2.2.3 Analysis

Before conducting any analysis, SAT scores, HS GPAs, assessment scores, test anxiety and self-efficacy scores were winsorized to two standard deviations from the mean (in order to maintain the direction of outliers while eliminating extreme values [108]). To determine if there were gender differences in pre and post TA, as well as assessment scores and prior academic preparation, we performed unpaired t-tests and calculated the Cohen's d between groups. Cohen's d is a measure of effect size, and we used the following standards: small, $d \sim 0.2$; medium, $d \sim 0.5$; and large, $d \sim 0.8$ [109].

Next, we conducted a mediation analysis of test anxiety and high-stakes assessment grades, using self-efficacy as a mediator. This model was chosen because self-efficacy had the largest Pearson correlation with test anxiety compared to all other variables. All involved variables were z-scored, meaning that observations were converted to measure the number of standard deviations they were from the mean so that regression weights can be directly compared without regard to their original units [108]. Mediation was conducted in R using the bootstrap method with 1000 simulations [110].

Finally, to explore the predictive relationships between test anxiety and assessment outcomes, we used multiple regression analysis. For each regression model, we report the standardized β coefficients, sample size, and R-squared. Standardized coefficients were used because they are in units of standard deviation and allow for direct comparison of effects [109]. We initially used gender, SAT math scores, and HS GPA as predictors for low- and high-stakes assessment scores. After establishing baseline models, we introduced pre TA or the average TA as predictor variables. Average test anxiety is the mean of pre and post test anxiety, and was used as a proxy for students' test anxiety while they were taking the course. During regression analysis, we used combined assessment categories (e.g., low and high-stakes assessments), but results were similar when the categories were separated. For example, the regression models predicting low-stakes assessment scores were similar to both the models predicting quiz grades and those predicting homework grades. All of the regression models predicting individual assessment types (i.e., homework, quizzes, midterm exams, and final exams) can be found in Tables 57 and 58 in Appendix A.

2.3 Results and Discussion

2.3.1 Are there gender differences in students' prior preparation, self-efficacy, or test anxiety?

Both men and women seemed well-prepared for an introductory STEM course, with SAT scores above the national average of 523 [111] and HS GPAs above 4.0, SAT math scores were significantly higher for men, while high school GPAs were non-significantly higher for women in this subset of students (see Table 29). Both HS GPA and SAT math scores have been shown to be predictors of undergraduate performance [112], especially in quantitative courses [113]. However, we cannot assume that men's higher SAT scores directly translate into physics performance, as students in this sample have no similar gap in Calculus 1, which they often take in tandem with Physics 1 [43–45].

Table 29 also shows that women reported higher levels of test anxiety and lower self-efficacy than men. For self-efficacy, the gender differences grew from small ($d \sim 0.2$) to medium ($d \sim 0.5$) over the semester; the test gender gap also increased, but maintained a large ($d \sim 0.8$) effect size. This is consistent with other studies that find gender differences in SE [40, 114] and TA [93] in the physics context. Growth in both TA and SE gender differences over the semester suggest that both constructs are malleable and that classroom experiences are affecting women's and men's test anxiety and self-efficacy differently.

2.3.2 Are there gender differences in students' low and high-stakes assessment scores?

From Table 29, homework and quiz grades show no significant gender difference, while midterm and final exams have significant medium ($d \sim 0.5$) differences. Although grade schemes differed by instructor, all courses had grades based primarily on midterm and final exam scores. Because research suggests that women are more likely than men to leave STEM fields due to concerns about grades (even if they have an A or B average) [115], women's lower exam scores may contribute to the loss of women from majors that require introductory calculus-based physics. This, combined with recent research that suggests that introductory mathematics courses are better predictors of future course success for physics and engineering students than introductory physics courses [43–45], suggests that many women who may have found success in advanced courses leave STEM fields before they have the opportunity to do so.

2.3.3 Are self-efficacy and test anxiety independent predictors of assessment scores?

The relationship between average self-efficacy and test anxiety is explored in a mediation model (see Figure 1). This model tests if average self-efficacy mediates the relationship between test anxiety and high stakes assessment scores, and was statistically significant. The Average Causal Mediation Effect (ACME) was -0.19 ($p < 0.001$), with a confidence interval of [-0.29,-0.09]. The average direct effect (ADE) was -0.24, $p = 0.006$ and the total direct effect (TDE) was -0.43, $p < 0.001$. Mediation models were similar for men and women when tested separately, so we combined them to maximize our sample size.

The model in Figure 1 suggests that test anxiety is partially mediated by self-efficacy. This means that test anxiety affects high-stakes assessment outcomes both directly and indirectly. One direct effect may be that test anxiety takes up students' cognitive resources during assessments [116]. In this model, the indirect effect refers to the effect of test-anxiety on self-efficacy (for example, a student thinking that they will never master physics because they always "freeze" when they're being assessed) and the subsequent effect of low self-efficacy on exam performance (that same student may decide that studying is not worth their time because they believe they will not receive a high grade even if they do put in the necessary effort).

Additionally, Bandura [63] theorized that poor performance and anxiety are "co-effects" of low self-efficacy. Heightened self-efficacy and reduced test anxiety likely form a virtuous cycle wherein students' development of coping mechanisms (for example, stress-reduction techniques and explicitly rehearsing strategies for academic challenges) increase their self-efficacy, and increased self-efficacy frees students' cognitive resources to focus on the task at hand and better implement coping strategies [63, 117].

Importantly, the relationship between test anxiety, self-efficacy and high-stakes assessments is similar for women and men. Prior work has shown similar results: anxiety affects academic self-efficacy similarly for men and women [118], but women tend to have lower self-efficacy than men regarding physics [11, 12, 40, 41, 65]. This is useful for instructors, as they do not have to focus on using different methods to aid students in developing self-efficacy or mitigating test anxiety, but rather using these methods more or less depending on the needs of the student.

2.3.4 Does gender predict scores in low- or high-stakes assessments when controlling for self-efficacy or test anxiety?

Tables 3 and 4 show the results of our eight regression models predicting low and high-stakes assessments, respectively. Each of the models uses the variables in the far left column to predict the outcome variable (e.g., low or high-stakes assessment scores). Any blank spaces in the table indicate that the predictor in the corresponding row was not used in that model. The strength of each predictor, controlling for all other predictors in the model, is given by the standardized regression (β) coefficient [109]. More specifically, for each change of one standard deviation in the predictor variable, the model predicts there will be a change of β standard deviations in the outcome variable, controlling for all other predictor variables [109].

In Models 1-4 (see Table 3), gender is not a significant predictor of low-stakes assessment scores, so there is no difference in motivational factors to account for. Only SAT Math and and pre SE were significant predictors of low-stakes assessment outcomes. Pre SE was only a predictor when pre TA was also included, as in Model 4 (as opposed to Model 2, which includes only TA, or Model 3, which includes only SE). However, SE is unlikely to have real-world effects on students' low-stakes assessment outcomes; all of the models in Table 3 explain only small amounts of the variance as R^2 never exceeds 0.06. Average TA and SE were not significant predictors (alone or together, see Table 58 in Appendix A).

Regression models predicting high-stakes assessments are found in Table 4. Model 5 shows that women have lower scores than men on high-stakes assessments when controlling for SAT math scores and HS GPA. Therefore, more factors are needed to account for this discrepancy. Pre SE and TA were not significant predictors, so models that included them did not account for the gender difference. However, including average TA (as in Model 6) or average SE (as in Model 7) renders the gender effect non-significant, suggesting that both TA and SE can explain some of the gendered performance differences in this class. If both average TA and SE are included, as in Model 8, TA becomes non-significant.

Prior work has found similar results: both TA and SE are significant, but TA does not explain more variance than SE alone, and TA becomes non-significant when SE is included [93]. These similarities suggest that differences in both test anxiety and self-efficacy are widespread in physics. However, the mediation model in Figure 1 and prior work which states that SE and TA are related [63], suggest that addressing test anxiety may be a way for instructors to aid their students in building self-efficacy. Thus, creating a low anxiety, equitable, and inclusive learning environment in which all students have a high sense of belonging and feel recognized by their instructors for their effort and progress is important, particularly for students from underrepresented groups who do not have role models, in order for them to master physics and develop confidence in their abilities to do so.

2.3.5 Teaching Recommendations

To help decrease student test anxiety and increase self-efficacy, educators should create an equitable and inclusive learning environment while maintaining high standards and provide students the tools and scaffolding support they need to become independent learners, utilizing frequent feedback [63]. The student sample in this study was 80% first-semester college students, who may be new to exams that make up the majority of their course grade. Implementing frequent, low-stakes assessment (for example, weekly or biweekly exams) can give students many attempts to practice test-taking and study skills and gives ample opportunity for feedback [119–121].

It also can space the practice (instead of cramming right before the exams, students will study more uniformly), which can lead to better retention of content and better skill development [122]. This can create a more equitable classroom environment by minimizing fears that come with test anxiety (e.g., receiving a low course grade due to one bad exam score) and also helps students develop the skills they will need if and when they encounter high-stakes exams in the future. In addition, instructors should decrease the importance of traditional exams in final course grades, while increasing the importance of clicker questions, homework, projects, and other assessments. A broad range of low-stakes formative assessments should be implemented throughout the course, to give students both frequent feedback and opportunities to master content [123, 124].

2.4 Summary and Conclusion

Our results show that there are gender differences in students' self-efficacy and test anxiety. When controlling only for prior preparation, there is also a gender gap in high-stakes assessment scores, which becomes non-significant when controlling for self efficacy and/or test anxiety. Lastly, self-efficacy mediates the relationship between test anxiety and high stakes assessment scores. From these findings we conclude that physics classrooms have the potential to become more equitable for women if instructors focus on giving students the support they need to enter a cycle of increasing self-efficacy and decreasing test anxiety. Further, we suggest that measuring student test anxiety in addition to self-efficacy in physics courses may be useful. If a student has low self-efficacy, there are a range of interventions and approaches that may improve their learning outcomes. However, students with test anxiety may have certain needs that are different from students who have not received encouragement from previous instructors, even though both groups of students may have low self-efficacy. By measuring and addressing the different mechanisms through which self-efficacy is developed, instructors and education researchers may be able to pinpoint the combination of supports students must be provided to excel.

This study mirrors results seen in other disciplines and institutions [93, 94]. As we collect more data, we want to include intersectional analysis to understand the relationship between gender, race, test anxiety, and self-efficacy. Additionally, it would be valuable to investigate how test anxiety and self-efficacy predict female and male student performance in low and high-stakes assessment in different countries in similar courses. It would also be valuable to investigate these issues in other contexts, e.g., at different types of institutions (e.g., large research university vs. small colleges). Outside of the physics context, studies about the relationship between text anxiety, self-efficacy, and low- and high-stakes assessment outcomes in other contexts (for example, other sciences or in the humanities) may be valuable to the larger education community. However, we hypothesize that test anxiety may be worse in physics than in other sciences because physics exams often have problems that are asked in very different contexts to the problems students solved earlier even though they involve similar underlying physics principles. For example, if students earlier learned that angular momentum conservation can be used to understand why a ballerina speeds up when she puts her arms close to herself, the exam question may ask them about the change in the angular speed of a white dwarf that is collapsing under its own gravitational force [125, 126]. Our prior research including individual interviews with students suggest that they struggle in transferring their learning from the ballerina context to the white dwarf context [125, 126] and it is particularly difficult for them when they are asked these types of isomorphic problems in timed exams in which they do not have sufficient time for contemplating that the underlying principles are similar for these problems. Although in interviews, some students mentioned that these types of physics problems in exams in which they have to transfer their learning from one context to another make them anxious, future studies would explicitly ask all students to rate their anxiety while answering such problems during high stakes and low stakes situations to investigate statistical differences.

Table 2: Means and standard deviations (SD) of prior preparation, test anxiety, self-efficacy, and assessment grades by gender. Cohen's d is negative when men have a higher score than women. * = p < 0.05, ** = p < 0.01, and *** = p < 0.001.

	Women		Men		
Variable (Range)	Mean	SD	Mean	SD	d
SAT Math (200-800)	691	55	718	51	-0.51**
HS GPA (0-4)	4.22	0.33	4.18	0.33	0.13
TA Pre (1-4)	2.83	1.02	2.11	0.79	0.82***
TA Post (1-4)	3.15	0.99	2.29	0.89	0.93***
SE Pre (1-4)	3.03	0.44	3.24	0.43	-0.49**
SE Post (1-4)	2.82	0.48	3.18	0.47	-0.75***
Homework (0-10)	9.16	0.52	9.04	0.59	0.19
Quiz (0-10)	9.36	0.58	9.48	0.46	-0.22
Midterm Avg. (0-10)	7.30	1.13	8.01	1.18	-0.61***
Final Exam (0-10)	5.50	1.73	6.41	1.75	-0.53***

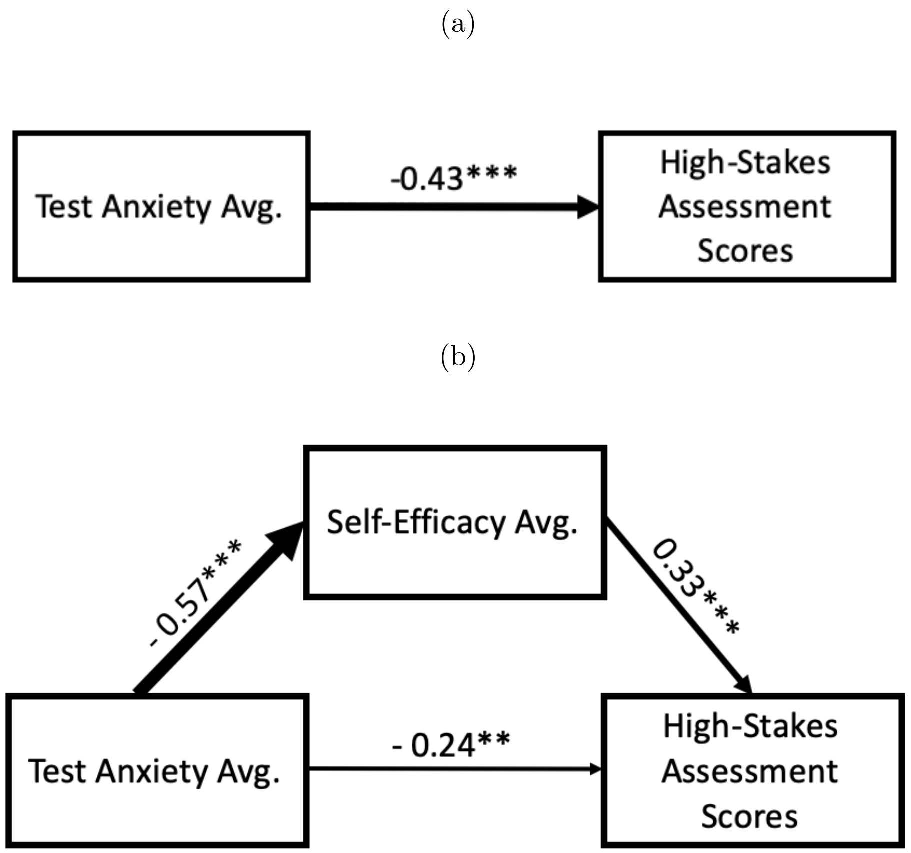


Figure 1: Mediation model results: (a) shows the model without self-efficacy, while (b) shows the model including self-efficacy. Average self-efficacy mediates the relationship between average test anxiety and high-stakes assessment scores. Line thickness corresponds to the standardized regression coefficient size. $N = 181$. $^{*} =$ p < 0.05, $^{**} =$ p < 0.01, and $^{***} =$ p < 0.001.

Table 3: Standardized β coefficients of regression models predicting low-stakes assessment scores. $N = 131$, “F” and “M” refer to female and male students. Significant predictors are bold. * = p < 0.05, ** = p < 0.01, and *** = p < 0.001.

Predictor:	Model 1	Model 2	Model 3	Model 4
Gender (F=1, M=0)	0.01	-0.03	0.04	0.00
SAT Math	**0.19***	**0.19***	**0.19***	**0.20***
HS GPA	0.11	0.11	0.12	0.12
Test Anxiety Pre		0.07		0.16
Self-Efficacy Pre			0.16	**0.22***
R^2	0.03	0.03	0.05	0.06

Table 4: Standardized β coefficients of regression models predicting high-stakes assessment scores. $N = 176$, “F” and “M” refer to female and male students. Significant predictors are bold. * = p < 0.05, ** = p < 0.01, and *** = p < 0.001.

Predictor:	Model 5	Model 6	Model 7	Model 8
Gender (F=1, M=0)	**-0.20****	-0.09	-0.08	-0.05
SAT Math	**0.46*****	**0.42*****	**0.43*****	**0.41*****
HS GPA	**0.23*****	**0.23*****	**0.25*****	**0.24*****
Test Anxiety Avg		**-0.26****		-0.10
Self-Efficacy Avg			**0.37****	**0.32****
R^2	0.38	0.43	0.49	0.49

3.0 Bioscience students with test anxiety have lower grades than those who don't: Low-stakes assessments could improve outcomes and increase gender equity in introductory physics courses

3.1 Introduction and theoretical framework

Students' science, technology, engineering, and mathematics (STEM) related motivational beliefs have implications for their performance in individual courses as well as their long-term outcomes [6–12]. In particular, students' motivational beliefs are correlated with their goals and tend to predict student learning outcomes [6–12]. There also tend to be differences between men's and women's motivational beliefs regarding physics, which have been linked to performance differences in physics courses [11, 16, 18, 42, 65]. Here, we focus on two motivational factors: test anxiety and self-efficacy. Test anxiety can affect students' test performance and is more likely to affect women [60]. Self-efficacy in a given domain is a student's belief in their ability to succeed at an activity or subject or complete a task [127].

Test anxiety can impact students' cognition, physical body, and behavior [60]. When they experience test anxiety, students' cognitive resources are not entirely devoted to the assessment, but can be taken up by worry and intrusive thoughts of failure [60]. Additionally, test anxiety can affect how students feel during an assessment. For example, they may experience a fast heartbeat or "butterflies in their stomach". The behavioral aspect of test anxiety manifests in avoidance techniques, such as procrastination or interacting only with surface-level feedback after the exam (e.g., not examining mistakes closely to make a plan for future improvement) [60,62]. Other studies have found that test anxiety negatively affects student performance, especially on high-stakes assessments such as exams [18,93,94]. In addition, women are more likely to report test anxiety than men [60, 127], so understanding test anxiety and how to minimize its effect on student success is vital to create inclusive and equitable learning environments.

Self-efficacy [64, 127] has been linked to positive learning outcomes for physics students [11,12,18,41,65]. Self-efficacy of students in a particular domain can be enhanced through several mechanisms. One way is by overcoming difficulties (such as a challenging homework assignment) [64]. Self efficacy can also be formed through social means, such as through seeing role models succeed in the domain of interest, and by encouragement to increase resolve and measure success via personal improvement [64]. The final mechanism is regulation of emotional states, such as management of anxiety [64].

Women commonly have lower physics self-efficacy than men [11, 12, 41, 65]. We hypothesize that one reason for this is that they tend to have fewer opportunities than men to develop self-efficacy. For example, they may have fewer experiences overcoming challenges in physics because they are less likely to take advanced physics courses in high school for a variety of reasons [66]. Additionally, women may have fewer role models due to under-representation of women in physics [66], and they are less likely to receive encouragement that they can succeed in physics from instructors and peers [28]. Because high self-efficacy allows students to develop coping mechanisms that could reduce test anxiety, we hypothesize that students with high self-efficacy are also likely to have low test anxiety [127].

In this research, we aim to investigate if test anxiety and/or self-efficacy can predict low- and high-stakes assessment outcomes. Here, low-stakes assessments are those that individually make up a small portion of a student's grade, such as homework. High-stakes assessments are individual assessments that make up a large portion of a student's grade such as traditional exams [18, 94]. For example, five assessments that each make up 10% of the student's grade are each lower-stakes than a single exam that makes up 50% of a student's grade, even if the content of each assessment is identical. Prior work has shown that gender gaps are more prevalent in high-stakes than low-stakes assessments [18, 94]. This, combined with gender differences in self-efficacy and test anxiety, leads us to hypothesize that test anxiety and self-efficacy may predict high-stakes, but not low-stakes assessment performance.

Past research shows a relationship between self-efficacy, test anxiety, and physics grades for students enrolled in introductory courses for engineering and physical science majors [18]. Additionally, prior research has found that there is a relationship between test anxiety and high-stakes assessment grades in biology classrooms [94]. Here, we focus on students in introductory physics for bioscience and heath-science related majors.

Students pursuing bioscience and health-science related majors are generally required to take at least one physics course for their major (and many of them are required to take two physics courses). Women are not underrepresented in these physics courses for bioscience and health-science related majors, but there may still be a gender gap in the motivational beliefs of students in the course. In particular, prior research has found that even in physics courses in which women are not underrepresented, men tend to have higher grades and physics-specific motivational beliefs than women [10, 16, 23, 33, 34, 128–135]. For example, women tend to have lower physics self-efficacy than men with the same grades in courses for engineering and physical science students as well as courses for students with interest in bioscience and health-science related professions [23, 41].

One goal of this research is to investigate the relationship between test anxiety and assessment outcomes, with a focus on gender differnces in each construct, for students majoring in bioscience and health-science related majors in introductory physics courses. We hypothesize that test anxiety will predict students' high-stakes, but not low-stakes assessment outcomes. We included self-efficacy in our investigation because the relationship between performance and self-efficacy is well-documented in physics [10, 41, 65], and because management of anxiety is explicitly mentioned as a mechanism to enhance self-efficacy [64]. With these goals in mind, we aim to answer the following research questions:

RQ1. Are there gender differences in students' prior preparation, self-efficacy, or test anxiety?

RQ2. Are there gender differences in students' low- or high-stakes assessment scores?

RQ3. Does gender predict performance on low- or high-stakes assessments when controlling for self-efficacy or test anxiety?

3.2 Methodology

3.2.1 Participants and Procedures

This study took place at a large research university in the United States. Participants were students enrolled in a Physics 1 or 2 course for bioscience and health-related majors. The Physics 1 course primarily covered mechanics, though both thermodynamics and waves were also included. The Physics 2 course covered electricity and magnetism, geometrical optics, and physical optics. Instructors taught the course in a traditional lecture-based format alongside smaller-sized recitations taught by teaching assistants in which students work collaboratively on physics problems.

The Physics 1 student sample included sections taught by two separate instructors, and the Physics 2 sample included sections taught by three separate instructors. Students were either given extra credit or a participation grade for taking the survey, depending on the instructor.

The surveys were given during the first and last week of their mandatory teaching assistant-led recitations. We call the first and final data sets ‘pre’ and ‘post’, respectively. In Physics 1, 426 students took the pre test and 422 took the post test. In Physics 2, 563 students took the pre test and 536 took the post test.

For analysis, we included only students who successfully passed an attention check on the survey (a question that requested the students select ‘C’). Additionally, we included as many students as possible in each part of the analysis. For example, in a model that uses the average of one construct as well students’ standardized test scores, we would exclude students who were missing SAT and ACT scores, or were missing either pre or post survey results. One Physics 1 class section and one Physics 2 class section was not able to fully complete the post survey and are missing post test anxiety data. This resulted in a smaller sample size for test anxiety post, but students in this section had statistically indistinguishable prior preparation, pre motivational factors, and assessment outcomes from other students in the sample, so they were included in analysis where possible.

This research was carried out in accordance with the principles outlined in this institution’s Institutional Review Board ethical policy, and de-identified demographic data were provided through university records. For some variables, such as high school GPA, this approach allows us to rely on records that may be more accurate than students’ memories. However, it limits other measures such as student sex/gender, which students could only report as “male” or “female”. We acknowledge the harm that collecting data this way can cause [136]. This institution recently began to implement more inclusive sex and gender reporting methods for students, which we plan to use once student samples are large enough to be meaningful in quantitative analysis. Demographic data indicated our Physics 1 sample was 66% women and our Physics 2 sample was 56% women. Students in Physics 1 identified with the following races/ethnicities: 62% White, 20% Asian, 4% Hispanic/Latinx, 6% multiracial, 8% African American/Black, and 1% unspecified. Students in Physics 2 identified with the following races/ethnicities: 60% White, 24% Asian, 4% Hispanic/Latinx, 5% multiracial, 5% African American/Black, and 1% unspecified.

3.2.2 Measures

3.2.2.1 Self-Efficacy and Test Anxiety

All test anxiety and self-efficacy survey items can be found in Table 5. The test anxiety survey questions were adapted from the previously validated Motivated Strategies for Learning Questionnaire [**?**,137]. To ensure we were measuring domain-specific mindset, we explicitly mentioned physics in the survey items, as seen in Table 5. For example 'I have an uneasy, upset feeling when I take an exam,' becomes 'I have an uneasy, upset feeling when I take a physics test'. Self-efficacy survey questions were constructed from other surveys and were previously validated [18, 23, 29]. Test anxiety items were either on a five-point Likert scale (1 - Not at all true, 2 - A little true, 3 - Somewhat true, 4 -Mostly true, 5 - Completely true) or a 7-point Likert scale (1 - Never true, 2 - Rarely true, 3 - Occasionally true, 4 - Neutral, 5 - Sometimes true, 6 -Usually true, 7 - Always true). Self-efficacy items were either on a four-point Likert scale (1-NO!, 2-no, 3-yes, 4-YES!) or a 7-point Likert scale (1-No!, 2-no, 3-Slightly leaning toward no, 4-Neutral, 5-Slightly leaning toward yes, 6-yes, 7-Yes!). All responses were placed on a 0-1 scale to account for multiple Likert scales.

Test Anxiety items were reverse coded so that a higher score indicates the student has low test anxiety or high self-efficacy. Thus, an ideal course outcome is that all students have low test anxiety and high self-efficacy scores. We further validated the survey through twenty one-hour student interviews to ensure that students interpreted questions as intended. Additionally, we performed confirmatory factor analysis using the students in this study as a check for continued validity.

For both the pre and post-surveys, the Comparative Fit Index (CFI) and Tucker Lewis Index (TLI) were ≥ 0.90 [106], the Root Mean Square Error of Approximation (RMSEA) and Standardized Root Mean Square Residual (SRMR) were both ≤ 0.08 [107], which can be seen in Appendix B in Table 59. Standardized factor loadings were all above 0.5 [106], which can be seen in Table 5. Cronbach's α was between 0.7 and 0.9 for all factors pre and post [108].

3.2.2.2 Prior academic preparation

High school Grade Point Average (HS GPA) was reported using the weighted 0–5 scale, which is based on the standard 0 (Failing)–4 (A) scale with adjustments for Honors, Advanced Placement and International Baccalaureate courses (all of these programs may offer a bonus of one or two grade points as a reward to taking advanced courses, which can allow a GPA higher than 4.0). High School GPA is taken as a measure of general academic skills and generally is a strong predictor of early undergraduate course performance [138].

Students' Scholastic Achievement Test math (SAT math) scores are on a scale of 200–800 and were used as a predictor of performance on high-stakes assessments involving mathematical problem-solving (e.g., physics exams) [113, 138, 139]. If a student took the American College Testing (ACT) examination, we converted ACT to SAT scores [102]. If a student took a test more than once, the school provided the highest section-level score for the SAT and the highest composite score for the ACT. If a student took both ACT and SAT tests, we used their SAT score.

3.2.2.3 Assessment Scores

Homework and exam grades were provided by instructors and were de-identified by an honest broker before being included in analysis. If grades were not on a 0-100 scale, they were rescaled. For example, if homework was graded on a 10-point scale, all scores were multiplied by 10 for analysis.

3.2.3 Analysis

First, we report means and standard deviations of each variable separately for men and women. Next, to determine if there were sex differences in the means of self-efficacy, test anxiety, prior preparation, or assessment scores we performed unpaired t-tests to measure statistical significance of the differences [108] and Cohen's d to measure the size of the difference [140]. Cohen's d is calculated using:

$$d = \frac{\mu_1 - \mu_2}{\sqrt{(\sigma_1^2 + \sigma_2^2)/2}},$$

where μ_1 and μ_2 are the mean values of each group and σ_1 and σ_2 are the standard deviations of each group [140]. Group one was women and group two was men. Cohen's d is considered small if $d \sim 0.2$, medium if $d \sim 0.5$, and large if $d \sim 0.8$ [140]. We performed this analysis separately for Physics 1 and Physics 2 courses.

To explore the predictive relationships between test anxiety and assessment outcomes, we used multiple regression analysis. For each regression model, we report the standardized β coefficients, sample size, and R-squared. Standardized coefficients were used because they are in units of standard deviation and allow for direct comparison of effects [141]. We initially used gender, SAT math scores, and HS GPA as predictors for low- and high-stakes assessment scores. Here, low-stakes assessment scores are the students' average homework grades. High-stakes assessment scores are weighted so that 75% of the category is midterm exam grades and 25% is the final exam grade. This weighting was done because the instructors gave three midterm exams and one final exam.

After establishing baseline models, we introduced pre or average test anxiety and self-efficacy as predictors. Average test anxiety/self-efficacy is the mean of pre and post scores, and was used as a proxy for students' test anxiety/self-efficacy while they were taking the course. During regression analysis, we used combined assessment categories (e.g., low and high-stakes assessments), but results were similar when the categories were separated. For example, the regression models predicting high-stakes assessment scores were similar to both the models predicting midterm exam grades and those predicting final exam grades.

3.3 Results and Discussion

3.3.1 RQ1. Are there gender differences in students' prior preparation, self-efficacy, or test anxiety?

For both Physics 1 and 2, men had higher SAT math scores than women, while women had higher high school GPAs than men. All these differences are small to medium ($d \sim 0.2$ to $d \sim 0.5$). Students were generally well-prepared for introductory physics: their average SAT Math score is well above the 2019 national average of 528 [**?**], and average high school GPA was around or above 4.0 on a 5.0 point scale. High school GPA and SAT Math scores are both predictors of undergraduate STEM performance [138]. However, we cannot assume that men's higher SAT scores directly translate into physics performance, as students in this sample show the reverse pattern in calculus 1, with women having a statistically significantly higher grade than men in this course [19]

Men tended to report higher self-efficacy and less test anxiety than women in both Physics 1 and Physics 2, for both the pre and post surveys. However, the magnitude of differences differed by construct and course. In Physics 1, gender differences were larger for test anxiety than for self-efficacy both pre and post. However, gender differences in text anxiety grew from pre to post, while they decreased in self-efficacy from pre to post. We note that the smaller gender gap in self efficacy at the end of the semester was the result of an average self-efficacy drop for both men and women from pre to post.

In Physics 2, gender differences were larger for test anxiety than for self-efficacy both on pre and post. Moreover, both test anxiety and self-efficacy gender gaps decreased from pre to post. However, this was the result of women's self-efficacy and test anxiety staying approximately the same, with very small drops in men's motivational factors over time.

In some ways, our results are similar to findings for students taking calculus-based introductory physics: there are gender differences in both self-efficacy and test anxiety favoring men [18], and the constructs change over time, showing that they are malleable and potentially able to be influenced [18]. However, this study does show some differences from prior work: it appears that in Physics 2 for students majoring in bioscience and health-science related majors, students' motivational factors do not decrease much over time.

We also note differences in motivational factors from Physics 1 to Physics 2. Students' self-efficacy decreased from Physics 1 pre to post, which can be seen in Figure 2. At the start of Physics 2, students reported self-efficacy that was similar to or slightly higher than what they reported at the start of Physics 1. From pre to post, self-efficacy in Physics 2 stayed fairly constant, which can also be seen in Figure 2. On the other hand, Figure 3 shows that students' average test anxiety gets worse over time, so that students reported the most test anxiety at the end of Physics 2. A student affected by test anxiety is likely to experience limits on the cognitive resources they can devote to the assessment [60], so we hypothesize that increased test anxiety may prevent students from accurately representing their knowledge on high-stakes assessments. In traditional exam-reliant courses, this is particularly concerning.

3.3.2 RQ2. Are there gender differences in students' low- and high-stakes assessment scores?

Homework constitutes the "low-stakes" assessment category. Tables 6 and 7 show that female students had higher homework scores than male students in Physics 1 (Table 6) and Physics 2 (Table 7). Both effect sizes were small ($d \sim 0.2$). Midterm and final exams constitute "high-stakes" assessments. In Physics 1, men tended to have higher exam scores. The gender difference was medium ($d \sim 0.5$) for midterm exams and small ($d \sim 0.2$) for final exams. On the other hand, there were no statistically significant gender differences in exam scores for Physics 2.

Although grade schemes differed by instructor, all courses had grades based primarily on midterm and final exam scores. This raises concerns for the Physics 1 course. Because research suggests that women are more likely than men to leave STEM fields due to concerns about grades (even if they have an A or B average) [115], women's lower exam scores may contribute to the loss of women from majors that require introductory physics. This, combined with data that shows that gender gaps exist in very few bioscience courses at this institution [19], suggests that many women who may have found success in advanced courses leave STEM fields before they have the opportunity to do so.

Physics 2 shows no such gender disparity in exam scores, though the course is graded in a similar way to Physics 1. Though this study is correlational in nature, one potential reason for this difference is the population (some bioscience and health science related majors only require students to take Physics 1). Another hypothesis is that students have had more time to develop coping mechanisms to mitigate the effect that factors such as test anxiety and self-efficacy have on exam performance.

3.3.3 RQ3. Does gender predict scores in low- or high-stakes assessments when controlling for self-efficacy or test anxiety?

For Physics 1, low-stakes assessment scores are not predicted by pre self-efficacy, average self-efficacy, pre test anxiety, or average test anxiety. The results of models predicting low-stakes assessments for Physics 1 and 2 can be found in Table 60 in Appendix B.

Physics 2 low-stakes assessment scores are not predicted by pre self-efficacy, average self-efficacy, or average test anxiety. Test anxiety scores negatively predict homework scores (i.e., students who report more test anxiety tend to have higher homework scores), but the effect size is small ($\beta = -0.10$, $p = 0.030$) and the model that includes test anxiety explains less of the variance than the model that excludes it (Adjusted $R^2 = 0.074$ versus 0.077). Because of the small effect size and drop in variance explained, this is unlikely to be a meaningful result. Results of models predicting low-stakes assessments for Physics 2 can be found in the Table 60 in Appendix B.

Tables 8 and 9 show the results of our regression models predicting high-stakes assessment outcomes. Each of the models uses the variables in the far left column to predict the outcome variable (e.g., low or high-stakes assessment scores). Any blank spaces in the table indicate that the predictor in the corresponding row was not used in that model. The strength of each predictor, controlling for all other predictors in the model, is given by the standardized regression coefficient [141]. More specifically, for each change of one standard deviation in the predictor variable, the model predicts there will be a change of beta standard deviations in the outcome variable, controlling for all other predictor variables [141].

For Physics 1, high stakes assessment scores are not predicted by pre test anxiety, but are predicted by pre self-efficacy, which can be seen in Table 8. However, Pre Model 1 in Table 8 (which includes self-efficacy and test anxiety as predictors) does not explain much more of the variance than Pre Model 2 (which does not include self-efficacy and test anxiety as predictors). Thus, pre self-efficacy does not appear to greatly contribute to performance differences among students. This is good because instructors can intervene during the course to improve students' self-efficacy and test anxiety.

However, in Physics 1, high-stakes assessment scores were predicted by average test anxiety but not self-efficacy, which can be seen in Table 8. Average Model 1 in Table 8 shows that average test anxiety predicts high-stakes assessment outcomes. Importantly, Average Model 1 explains more of the variance compared to Average Model 2. Additionally, there are no statistically significant gender differences in Average Model 1 which includes test anxiety as a predictor of high-stakes grades, but there are in Average Model 2. This means that gender differences in test anxiety may account for at least some of the gender discrepancies we see in high-stakes assessments.

For Physics 2, high-stakes assessment scores are not predicted by either pre self-efficacy or pre test anxiety. However, both average self-efficacy and average test anxiety predict high-stakes assessment scores, which can be seen in Table 9.

Broadly, we find that self-efficacy positively predicts high-stakes assessment scores, while test anxiety negatively predicts scores. We also found in Section 3.3.1.RQ1 that both self-efficacy and test anxiety measures became worse over time for all students, and that women reported lower self-efficacy and more test anxiety than men. Additionally, women in Physics 1 had lower high-stakes assessment scores than men. Thus, it is important to take steps to reduce student test anxiety and increase student self-efficacy. This is important to encourage the success of all students, but particularly women who appear to be more affected by low self-efficacy and high test anxiety, especially in Physics 1.

Test anxiety and self-efficacy can both be improved through providing coping strategies for students. For example, frequent assessment gives students many attempts to practice test-taking and encourages spaced practice which is more effective for retention and skill development than “cramming” before an exam [122]. Additionally, implementing a range of assessments (such as clicker questions, homework, tutorials, projects, and other types of assessments), each of which do not count for a very large portion of a students’ course grade, can help students develop a wider variety of skills without increasing anxiety. Providing students with these supports can help students develop the skills they will need if and when they encounter high-stakes exams in the future. Additionally, instructors can help decrease test anxiety by directly decreasing the importance of high-stakes assessments and increasing the importance of low-stakes assessments in their course. This can create a more equitable classroom environment by minimizing fears that come with test anxiety (e.g., receiving a low course grade due to one bad exam score).

3.4 Conclusions

In summary, test anxiety and self-efficacy predict high-stakes assessment outcomes. Additionally, women tend to have worse outcomes for self-efficacy, test anxiety, and high-stakes assessment outcomes than men in an introductory physics course for bioscience and health science related majors. Finally, we note that students’ self-efficacy and test anxiety tended to get worse from the start to the end of the semester: this is a poor outcome for all students and is particularly detrimental to women in the course. To help decrease student test anxiety and increase self-efficacy, educators should create an equitable and inclusive learning environment while maintaining high standards and provide students the tools and scaffolding support they need to become independent learners, utilizing frequent feedback.

Table 5: Items included in student survey. Items 5-8 were reverse coded. The same items were given to students for the pre and post survey. Students were included in factor analysis if they competed the pre or post survey.

		Factor Loading			
		Physics 1		Physice 2	
	Construct Name/Item Text	Pre	Post	Pre	Post
	Self-Efficacy				
1.	I am able to help my classmates with physics in the laboratory or in recitation	0.57	0.54	0.58	0.57
2.	I understand concepts I have studied in physics	0.58	0.73	0.70	0.69
3.	If I study, I will do well on a physics test	0.71	0.82	0.85	0.84
4.	If I encounter a setback in a physics exam, I can overcome it	0.73	0.80	0.84	0.89
	Test Anxiety				
5.	I am so nervous during a physics test that I cannot remember what I have learned	0.85	0.81	0.85	0.86
6.	I have an uneasy, upset feeling when I take a physics test	0.90	0.92	0.93	0.91
7.	I worry a great deal about physics tests	0.83	0.87	0.86	0.84
8.	When I take a physics test, I think about how poorly I am doing	0.83	0.80	0.85	0.86

Table 6: Sample size, mean, standard deviation (SD), and comparison of prior preparation, motivational factors, and assessment outcomes of students enrolled in Physics 1 during remote and in-person instruction. Results and significance of unpaired t-tests are provided. Cohen's d effect sizes are also given; a negative d indicates that female students had lower scores than male students.

	Female			Male			Comparison		
Variable	N	Mean	SD	N	Mean	SD	t	p	d
HS GPA	321	4.17	0.40	162	3.98	0.54	4.25	<0.001	0.41
SAT Math	310	668	69	158	684	70	-2.34	0.020	-0.23
Self-Efficacy Pre	285	0.60	0.16	139	0.69	0.14	-5.43	<0.001	-0.56
Self-Efficacy Post	280	0.53	0.19	140	0.62	0.20	-4.38	<0.001	-0.45
Test Anxiety Pre	277	0.50	0.25	138	0.68	0.25	-6.89	<0.001	-0.72
Test Anxiety Post	142	0.40	0.26	68	0.62	0.25	-5.63	<0.001	-0.49
Homework	321	93	11	162	90	17	2.32	0.21	0.22
Midterm Exams	321	67	16	162	73	15	-4.40	<0.001	-0.42
Final Exam	321	61	17	162	65	16	-2.41	0.016	-0.23

Table 7: Sample size, mean, standard deviation (SD), and comparison of prior preparation, motivational factors, and assessment outcomes of students enrolled in Physics 2 during remote and in-person instruction. Results and significance of unpaired t-tests are provided. Cohen's d effect sizes are also given; a negative d indicates that female students had lower scores than male students.

	Female			Male			Comparison		
Variable	N	Mean	SD	N	Mean	SD	t	p	d
HS GPA	350	4.20	0.38	276	4.07	0.44	4.07	<0.001	0.33
SAT Math	346	679	71	272	699	66	-3.58	<0.001	-0.29
Self-Efficacy Pre	301	0.62	0.16	240	0.70	0.16	-5.88	<0.001	-0.51
Self-Efficacy Post	299	0.60	0.18	208	0.67	0.17	-4.44	<0.001	-0.40
Test Anxiety Pre	314	0.37	0.24	244	0.54	0.27	-8.07	<0.001	-0.69
Test Anxiety Post	223	0.38	0.25	143	0.51	0.26	-4.61	<0.001	-0.49
Homework	350	95	11	276	91	19	3.48	<0.001	0.28
Midterm Exams	350	76	18	276	76	20	-0.34	0.738	-0.03
Final Exam	350	69	21	276	67	25	0.97	0.334	0.08

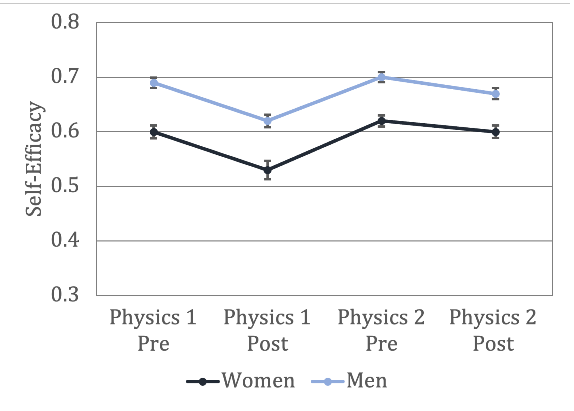


Figure 2: Average self-efficacy scores of men and women from the start of Physics 1 to the end of Physics 2. Error bars represent standard error and self-efficacy is on a 0-1 scale.

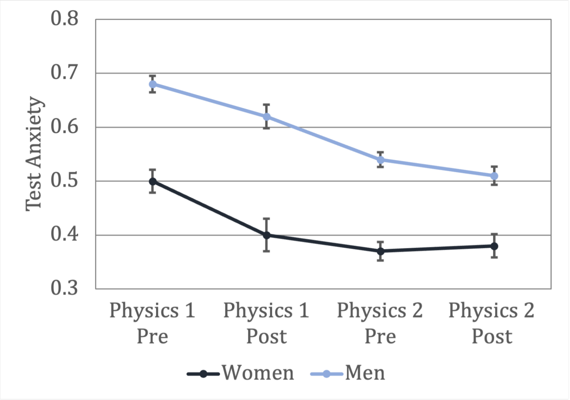


Figure 3: Average test anxiety scores of men and women from the start of Physics 1 to the end of Physics 2. Error bars represent standard error and test anxiety is on a 0-1 scale.

Table 8: Physics 1 high-stakes assessment scores predicted by student sex, High School GPA (HS GPA), SAT/ACT Math scores, average self-efficacy and average test anxiety. Standardized regression (β) coefficients are provided. $^{*} = p < 0.05$, $^{**} = p < 0.01$, $^{***} = p < 0.001$, and ^{ns} = not statistically significant.

	Pre		Average	
Variable	Model 1	Model 2	Model 1	Model 2
Sex	-0.16***	-0.19***	-0.06^{ns}	-0.20**
HS GPA	0.23***	0.23***	0.23***	0.29***
SAT/ACT Math	0.46***	0.48***	0.38***	0.43***
Self-Efficacy	0.13**		0.12^{ns}	
Test Anxiety	-0.01^{ns}		0.26***	
Adjusted R^2	0.38	0.37	0.45	0.37
N	399	399	174	174

Table 9: Physics 2 high-stakes assessment scores predicted by student sex, High School GPA (HS GPA), SAT/ACT Math scores, average self-efficacy and average test anxiety. Standardized regression (β) coefficients are provided. $^{*} = p < 0.05$, $^{**} = p < 0.01$, $^{***} = p < 0.001$, and ^{ns} = not statistically significant.

	Pre		Average	
Variable	Model 1	Model 2	Model 1	Model 2
Sex	0.01^{ns}	-0.01^{ns}	-0.03^{ns}	-0.10^{ns}
HS GPA	0.31***	0.31***	0.23***	0.23***
SAT/ACT Math	0.28***	0.29***	0.30***	0.35***
midrule Self-Efficacy	0.07^{ns}		0.16**	
Test Anxiety	0.01^{ns}		0.14**	
Adjusted R^2	0.22	0.22	0.27	0.22
N	304	304	533	533

4.0 Introductory physics during COVID-19 remote instruction: Gender gaps in exams are eliminated, but self-efficacy and test anxiety still predict success

4.1 Introduction

Assessing student learning outcomes and improving diversity in physics departments and classrooms during the COVID-19 pandemic are issues in physics education that many researchers are trying to address, especially regarding differences between online and in-person courses [142–145]. Some studies have found that remote instruction during COVID-19 correlates with lower student motivation and performance [142]. However, some researchers found that there was no significant drop in student learning outcomes or motivation during the COVID-19 pandemic [143]. Other studies have found that prior physics knowledge and homework completion are both stronger predictors of exam grades than whether the class was online or in-person [144].

Prior research suggests that student performance on content-based surveys does not significantly differ between in-person and online administration [146]. Some studies have also found that answer copying on homework problems is not more prevalent during remote than in-person instruction [147]. Because of this, we make the assumption that assessment outcome differences between in-person and online courses are not inherent, but may be the result of instructor choices in class policies. For example, at this institution, online and in-person physics courses had different exam policies. During online instruction, students were given a two part exam in which a group exam was followed by an individual exam. This approach decreases the importance of each individual assessment because they were worth a smaller portion of their overall grade (i.e., the group and individual sections combined were the worth the same amount of points as one traditional exam during in-person classes). Additionally, prior studies have fund that students have better individual assessment outcomes after working in a group [148]. Researchers suggest that this is partly due to co-construction of knowledge: that is, students are able to combine knowledge to correctly answer questions neither one would be able to answer on their own [148]. Because some prior research shows that students tend to have higher grades for online than in-person courses [149], and due to different grading approaches for the two course types, we anticipate that remote classes will have higher assessment scores than in-person courses.

Motivational factors tend to predict learning outcomes both inside and outside of physics classrooms [6–8, 150], so in this study we compare two motivational factors during in-person versus online classes: self-efficacy and test anxiety. Self-efficacy is someone's belief in their capability to succeed at an activity or subject [63, 64], and has been linked to positive outcomes for physics students [11, 12, 41, 42, 65]. Self efficacy can be developed in a variety of ways. These mechanisms include experiences successfully overcoming obstacles, seeing peers or relatable role models succeeding, and interpersonal encouragement that one can overcome challenges in a domain [64]. A fourth mechanism that is less commonly studied is management of physical and emotional states [64]. One such emotional state is test anxiety [63].

Test anxiety is a phenomenon that affects students in a variety of ways. One such way is physiological: a student may experience "butterflies in their stomach" or a fast heartbeat [95]. Another example is worry and self-preoccupation (e.g., thinking about one's perceived shortcomings instead of the task at hand) and intrusive thoughts of failure. A student affected by test anxiety in this manner will experience limits on time and cognitive resources they can devote to the assessment [95]. A third way test anxiety may affect students in how they prepare for and reflect back on exams. Students may procrastinate or choose to avoid looking at assessment feedback once it is graded [95, 101]. Managing test anxiety can help students increase their self-efficacy, and high self-efficacy can help students manage test anxiety [63, 151]. Instructors may be able to take advantage of this "virtuous cycle" of increasing self-efficacy and decreasing test anxiety to improve students motivation and learning outcomes.

Because self-efficacy allows students to develop coping mechanisms that could thwart test anxiety [63], we hypothesize that students with high self-efficacy will also have low test anxiety. To explore how test anxiety and self-efficacy predict students' performance in different situations, we first compare the performance of students taking in-person and online classes. We compared outcomes separately for low-stakes (for example, homework and quizzes) and high-stakes (for example, exams) assessments. In this paradigm, ten assessments that each make up five percent of a student's grade would be relatively low-stakes, while a single assessment that makes up 50 percent of a student's grade would be high-stakes, even if the content was identical between the assessments. Prior research has found that test anxiety and self-efficacy both predict high-stakes; but not low-stakes assessment outcomes [18]. Thus, we hypothesize that a course structure with more assessments will trigger less test anxiety and students will have higher grades. In this case, this suggests students will report less test anxiety during online classes because exams were given in two parts: group followed by individual. Additionally, we predict that, because the management of anxiety can contribute to high self-efficacy (and vice versa) [63], test-anxiety may predict self-efficacy.

In addition to comparing students' mean low- and high-stakes assessment scores between in-person and remote classes, we also compare gender differences in each assessment category. Specifically, we compare female and male students' performance on low-stakes and high-stakes assessments for both online and in-person classes. Gendered performance differences in introductory physics and other science courses have been found by many researchers [11,12,22,32,41,42,65,93,152], and these performance differences are commonly attributed to differences in motivational beliefs. These differences can be at least partially attributed to societal stereotypes about who can excel in physics [11,12,41,42,65,93,94]. Specifically, test anxiety [18,93,94] and self-efficacy [11,12,42,65], among other factors, have been linked to gender differences in performance.

Female students tend to have lower self-efficacy than male students in physics, even if they earn the same grade in the course [11,12,41,42,65]. gender differences in self-efficacy may be due to inequitable opportunities for women in physics. One way students form self-efficacy is through mastery experiences. Because women are less likely to take advanced physics courses in high school, they are less likely to have previous mastery experiences [66]. Women also tend to have fewer role models in physics due to the underrepresentation of women in the field [66]. Though it is not necessary for a potential role model to share all of a student's identities, prior work has shown that women especially benefit from other women as peers and role models [98,99]. Additionally, women, in general, are more likely to report test anxiety than men [60,63], so understanding test anxiety may play a key role in creating equitable learning environments.

Because previous studies have shown larger gender gaps in high-stakes than low-stakes assessment [94], and women are more likely to report low self-efficacy [41] and high test anxiety [60,63] than men, we hypothesize that test anxiety and self-efficacy will more strongly predict high-stakes than low-stakes assessment outcomes.

Additionally, there is some evidence that online learning particularly benefits female students [153]. Thus, we also predict that the course type with the lowest-stake assessments (for example, more exams or exams in multiple parts) will have smaller gender differences in performance.

Broadly, we are interested in how in-person courses compare to remote classes. We are interested in overall learning outcome differences, as well as differences in gender equity three times: instruction prior to the COVID-19 pandemic ("pre-remote"), remote instruction due to the COVID-19 pandemic ("remote"), and in-person instruction after the two semesters of remote instuction ("post-remote"). More specifically, we aim to answer the following research questions in the context course taught during the three time periods:

RQ1. Overall differences between in-person versus remote instruction

  a. How do the means of students' self-efficacy and test anxiety differ during remote versus in-person instruction?

  b. How do the means and distributions of students' high school GPAs, SAT/ACT math scores, and low- and high-stakes assessment outcomes differ during remote versus in-person instruction?

RQ2. gender differences during in-person versus remote instruction

  a. How do gender differences in students' self-efficacy and test anxiety differ during remote versus in-person instruction?

  b. How do gender differences in students' high school GPAs, SAT/ACT math scores, low- and high-stakes assessment outcomes differ during remote versus in-person instruction?

RQ 3. Predicting assessment outcomes during in-person versus remote instruction

a. Which factors predict low-stakes assessment scores during remote and in-person instruction?

b. Which factors predict high-stakes assessment scores during remote and in-person instruction?

4.2 Methods

4.2.1 Participants and Procedures

This study took place at a large research university in the United States. We administered a multiple-choice motivational survey to students in introductory calculus-based Physics 1 and 2 courses. The surveys were given during the first and last week of their teaching assistant-led recitations. We call the first and final data sets "pre" and "post", respectively. For analysis, we included only students who successfully passed an attention check (a question that requested the students select 'C') on both. Students were given extra credit for taking the survey.

This research was carried out in accordance with the principles outlined in this institution's Institutional Review Board ethical policy, and de-identified demographic data were provided through university records. For some variables, such as high school GPA, this approach allows us to rely on records that may be more accurate than students' memories. However, it limits other measures such as student gender. Historically, this institution only allowed students to select "male" or "female" for their gender. We acknowledge the harm that collecting data this way using a label that conflates gender and sex can cause [135, 136]. In this paper, we refer to this variable as "gender", although "male" and "female" are used in some places consistent with the label used by the institution. This institution recently began to collect separate information on student sex and gender, which we plan to use once student samples are large enough to be meaningful in quantitative analysis.

There was a total of 838 enrolled students in Physics 1, 87% of students took the pre survey, 73% took the post survey, and 67% took both. One class with 191 students was not able to complete post test anxiety questions. However, the grades and demographics were indistinguishable from students who were able to complete the post test anxiety questions. Demographic data indicated our sample was 37% women. Students identified with the following races/ethnicities: 70% White, 14% Asian, 6% Hispanic/Latinx, 4% multiracial, 4% African American/Black, and 2% unspecified.

There was a total of 603 enrolled students in Physics 2, 77% of students took the pre survey, 65% took the post survey, and 52% took both. Demographic data indicates our sample was 37% female students. Students identified with the following races/ethnicities: 66% White, 18% Asian, 4% Hispanic/Latinx, 5% multiracial, 3% African American/Black, and 2% unspecified. For both Physics 1 and 2, the majority of students are physical science or engineering majors in their first year of university.

Less that 15% of students at this institution received a Pell grant and 34% of students have a household income $\leq$ 200% of the federal poverty level. Additionally, at this institution, 6-year graduation rates are above 80% for all races and genders of students, regardless of income level [46]. Thus, we assume that assessment outcome differences between remote and in-person classes were primarily due to differences in class structure rather than a direct result of student difficulties outside of their classes during the pandemic when classes were moved online.

4.2.2 Course description

All courses included in this analysis were taught by one instructor with over a decade of experience regularly teaching introductory physics sequence courses for engineering and physical science majors. Though the course was primarily lecture-based, instruction included some pre-class online lectures and active learning approaches, such as clicker questions and in-class problem solving. Because this class has both synchronous and asynchronous activities/lectures when offered both in-person and remotely, these courses may be more comparable than other instances of in-person versus remote teaching. Physics 1 covers mechanics and waves, while Physics 2 covers electricity, magnetism, circuits, and physical optics.

remote classes were held for two semesters as a result of the COVID-19 pandemic, but this institution typically offers only in-person classes in physics. For both Physics 1 and 2, in-person and remote classes had a grading scheme in which the final exam made up 20% of final grades, three midterm exams that made up 40%, recitation quizzes that made up 10%, homework that made up 10%, and various other in-class assignments (such as clicker questions and open-book concept quizzes to check that students kept up with readings) that made up 20%.

For in-person classes, quizzes were completed in groups and exams were completed individually. For remote classes, quizzes were also completed in groups. However, exams (both midterm and final) during remote classes consisted of two parts. In the first part, students completed two questions in groups, followed by four questions individually. The group and individual problems each made up 50% of the exam grade.

4.2.3 Measures

4.2.3.1 Self-Efficacy and Test Anxiety

All test anxiety and self-efficacy survey items can be found in Table 10. The test anxiety survey questions were adapted from the previously validated [104] Motivated Strategies for Learning Questionnaire [105]. To ensure we were measuring domain-specific mindset, we explicitly mentioned physics in the survey items, as seen in Table 10. For example 'I have an uneasy, upset feeling when I take an exam,' becomes 'I have an uneasy, upset feeling when I take a physics test'. Self-efficacy survey questions were constructed from other surveys and previously validated [41, 42].

Test anxiety items were either on a five-point Likert scale (1-Not at all true, 2-S little true, 3-Somewhat true, 4-Mostly true, 5-Completely true) or a 7-point Likert scale (1- Never true, 2-Rarely true, 3-Occasionally true, 4-Neutral, 5-Sometimes true, 6-Usually true, 7-Always true). Self-efficacy items were either on a four-point Likert scale (1-NO!, 2-no, 3-yes, 4-YES!) or a 7-point Likert scale (1-No!, 2-no, 3-Slightly leaning toward no, 4-Neutral, 5-Slightly leaning toward yes, 6-yes, 7-Yes!). All responses were placed on a 0-1 scale to account for multiple Likert scales. Test Anxiety items were reverse coded so that a higher score indicates the student has low test anxiety or high self-efficacy. Thus, an ideal course outcome is that all students have low test anxiety and high self-efficacy scores.

We further validated the survey through twenty one-hour student interviews to ensure that students interpreted questions as intended. Additionally, we performed confirmatory factor analysis using the students in this study as a check for continued validity. For both the pre and post-surveys, the Comparative Fit Index (CFI) and Tucker Lewis Index (TLI) were ≥ 0.95 [106], the Root Mean Square Error of

Table 10: Items included in student survey. Items 5-8 were reverse coded. The same items were given to student for the pre and post survey.

	Self-Efficacy
1.	I am able to help my classmates with physics in the laboratory or in recitation
2.	I understand concepts I have studied in physics
3.	If I study, I will do well on a physics test
4.	If I encounter a setback in a physics exam, I can overcome it
	Test Anxiety
5.	I am so nervous during a physics test that I cannot remember what I have learned
6.	I have an uneasy, upset feeling when I take a physics test
7.	I worry a great deal about physics tests
8.	When I take a physics test, I think about how poorly I am doing

Approximation (RMSEA) and Standardized Root Mean Square Residual (SRMR) were both ≤ 0.08 [107], which can be seen Table 61 of Appendix C. Standardized factor loadings were all above 0.5 [106], which can be see in Table 62. Cronbach's α was between 0.7 and 0.9 for all factors pre and post [108].

4.2.3.2 Prior academic preparation

High school Grade Point Average (HS GPA) was reported using the weighted 0–5 scale, which is based on the standard 0 (Failing)–4 (A) scale with adjustments for Honors, Advanced Placement and International Baccalaureate courses. High School GPA is taken as a measure of general academic skills and generally is a strong predictor of early undergraduate course performance [138]

Students' Scholastic Achievement Test math (SAT math) scores are on a scale of 200–800 and were used as predictor of performance on high-stakes assessments involving mathematical problem-solving (e.g., physics exams) [113, 138, 139]. If a student took the American College Testing (ACT) examination, we converted ACT to SAT scores [102]. If a student took a test more than once the school provided the highest section-level score for the SAT and the highest composite score for the ACT. If a student took both the ACT and SAT tests, we used their SAT score. Some students (97 taking Physics 1 and 57 taking Physics 2) did not take the SAT or ACT because this institution became test-optional during the COVID-19 pandemic.

4.2.3.3 Course Grade

Course grades were based on the 0-4 scale used at our university, with A = 4, B = 3, C = 2, D = 1, F = 0 or W (late withdrawal), where the suffixes '+' and '-', respectively, add or subtract 0.25 grade points (e.g., B- = 2.75 and B+ = 3.25), except for the A+, which is reported as 4.

4.2.4 Analysis

To determine if there were mean differences in self-efficacy, test anxiety, prior preparation, and assessment scores between remote and in-person classes, we performed unpaired t-tests and calculated Cohen's d between groups. Cohen's d is a measure of effect size that is calculated using:

$$d = \frac{\mu_1 - \mu_2}{\sqrt{(\sigma_1^2 + \sigma_2^2)/2}}, \tag{1}$$

where μ_1 and μ_2 are the mean values for each group and σ_1 and σ_2 are the standard deviations for each group [140]. For remote versus in-person analysis, group one was in-person and group two was remote. For sex-based analysis, group one was female students and group two was male students. We used the following standards for effect size: small, d $\sim$ 0.2; medium, d $\sim$ 0.5; and large, d $\sim$ 0.8 [109]. We performed this analysis separately for Physics 1 and Physics 2 courses. Additionally, we created histograms (using Stata [154]) of the distributions of each variable for both in-person and remote classes. Next, to determine if there were sex differences in the means of self-efficacy, test anxiety, prior preparation, and assessment scores for both remote and in-person classes, we performed unpaired t-tests and calculated Cohen's d between male and female students. Again, this analysis was performed separately for Physics 1 and 2.

To explore the predictive relationships between test anxiety and assessment outcomes, we used multiple regression analysis. For each regression model, we report the standardized β coefficients, sample size, and Adjusted R-squared. Standardized coefficients were used because they are in units of standard deviation and allow for direct comparison of effects [109]. We initially used student sex, SAT math scores, high school GPA, test anxiety, and self-efficacy as predictors for low- and high-stakes assessment scores. Here, low-stakes assessment scores are the average of homework and quiz grades. High-stakes assessment scores are weighted so that 75% of the category is midterm exam grades and 25% is the final exam grade. This weighting was done because the instructor gave three midterm exams and one final exam.

Our first model used student sex, SAT/ACT math scores, high school GPA, test anxiety, and self-efficacy as predictors of assessment outcomes. Our second model removed test anxiety, leaving only student sex, SAT/ACT math scores, high school GPA, and self-efficacy as predictors. Our third model removed self-efficacy but included test anxiety. Our fourth and final model only used student sex, SAT/ACT math scores, and high school GPA, and acted as a baseline model.

Average test anxiety is the mean of pre and post test anxiety, and was used as a proxy for students' test anxiety while they were taking the course. Average self-efficacy was found the same way. During regression analysis, we used combined assessment categories (e.g., low and high-stakes assessments), but results were similar when the categories were separated. For example, the regression models predicting low-stakes assessment scores were similar to both the models predicting quiz grades and those predicting homework grades.

As we conducted regression analysis, we also used mediation where appropriate [155, p. 49-50]. First, test anxiety predicts assessment outcomes in a linear regression model. Second, test anxiety predicts self-efficacy in a linear regression model. And third, test anxiety becomes a weaker predictor of assessment outcomes if it is included with self-efficacy in a multiple regression model [155, p. 49-50]. For both mediation and regression analysis, all involved variables were z-scored, meaning that observations were converted to measure the number of standard deviations they were from the mean so that regression weights can be directly compared without regard to their original units [108]. Mediation was conducted in R using the bootstrap method with 1000 simulations [110].

4.3 Results

4.3.1 Overall differences between in-person versus remote instruction

4.3.1.1 How do the means of students' self-efficacy and test anxiety differ prior to remote versus in-person instruction?

Mean self-efficacy and test anxiety scores were indistinguishable between students enrolled in remote versus in-person Physics 1 classes. Table 11 shows that there was no statistically significant difference in pre or post self-efficacy or test anxiety scores between students enrolled in remote and in-person classes.

Unlike students enrolled in Physics 1, students enrolled in Physics 2 had some differences in motivational factors between the in-person and remote groups. Table 12 reveals that there is a small ($d \sim 0.2$) difference in pre self-efficacy scores: students taking the in-person class entered the course with slightly higher self-efficacy than those taking the remote class. However, differences in self-efficacy became non-significant by the end of the course. Students taking Physics 2 entered with similar test anxiety at the beginning of the course, but by the end students enrolled in remote classes reported more test anxiety then those taking in-person classes, though the difference was small ($d \sim 0.2$).

4.3.1.2 How do the means and distributions of students' high school GPAs, SAT/ACT math scores, and low- and high-stakes assessment outcomes differ during remote versus in-person instruction?

Students taking Physics 1 online tended to have slightly higher high school GPAs than those who took Physics 1 in-person, though Table 11 shows that the difference was small ($d \sim 0.2$). Physics 1 students taking in person and remote classes had statistically indistinguishable SAT/ACT math scores. Students taking Physics 2 remotely and in-person had statistically indistinguishable high school GPAs and SAT/ACT math scores, which can be seen in Table 12.

There was no statistically significant difference in homework grades between in-person and remote instruction for students enrolled in either Physics 1 or Physics 2 (see Tables 11 and 12). For students enrolled in Physics 1, quiz grades were higher during remote instruction, and Table 11 shows that this difference had a moderate effect size ($d \sim 0.5$). There was no statistically significant difference between quiz grades for remote and in-person instruction for students enrolled, shown in Table 12.

In both Physics 1 and Physics 2 courses, high-stakes assessments grades were higher for students during remote than in-person instruction. The differences were more pronounced in Physics 2 than Physics 1. For example, Table 11 shows that the grade differences in Physics 1 were moderate ($d \sim 0.5$) for midterm exams and large ($d \sim 0.8$) for final exams. However, Table 12 shows that the grade differences in Physics 2 were large ($d \sim 0.8$) for midterm exams and extremely large ($d_{final} = -1.71$) for final exams.

Notably, the difference in final exam scores between remote and in-person instruction were larger than for any other assessment category. Figure 8 shows that the distribution of final exam scores is both higher and more narrowly distributed for remote than for in-person classes. This figure also shows that very few students received a final exam lower than 50% during remote Physics 1 or Physics 2 courses, while a score of 50% is within one standard deviation of the mean final exam score for both Physics 1 and 2 courses taken in-person.

4.3.2 Gender differences during in-person versus remote instruction

4.3.2.1 How do gender differences in students' self-efficacy and test anxiety differ during remote versus in-person instruction?

Men tended to report higher self-efficacy and less test anxiety than women in both Physics 1 and Physics 2, for both the pre and post surveys. However, the magnitude of differences differed by construct. Generally, students entered both in-person and online classes with similar self-efficacy gender differences, but those differences evolved differently over time. In Physics 1, students entered both remote and in-person classes with medium sex difference effect sizes ($d \sim 0.5$), which can be seen in Table 13. This difference decreased more for remote classes than for in-person classes: by the end of the semester, the gender difference was still medium for in-person classes, but the gender difference was small ($d \sim 0.2$) and not statistically significant for remote classes. It should be noted that, although gender differences decrease over time, mean motivational scores for all students stayed approximately the same or decreased from pre to post.

At the start of Physics 2, Table 14 shows that students entered in-person and remote classes with identical self-efficacy gender differences (though students in remote classes entered with lower self-efficacy than those in in-person classes). This gender difference decreased during in-person classes, but increased during remote classes. Broadly, pre self-efficacy was similar regardless of whether the class was remote or in-person. Self-efficacy gender differences decreased from pre to post in Physics 1, regardless of whether the class was remote or in-person. However, self-efficacy gender differences decreased for Physics 2 in-person classes and increased for Physics 2 remote classes.

In Physics 1, text anxiety gender differences decreased over time for both in-person and remote courses, though this difference was larger at the start of the semester for in-person classes, which can be seen in Table 13. For Physics 2, gender differences increased over time during in-person classes, but remained stagnant during online classes, as seen in Table 14.

In conclusion, most gender differences in test anxiety and self-efficacy decrease or remain the same from the start to the end of the semester. However, in-person Physics 2 classes had an increased gender difference in test anxiety from the start to the end of the semester, and remote Physics 2 courses had an increased difference in self-efficacy from the start to the end of the semester. From an equity standpoint, gender differences decreased more during remote than in-person Physics 1 classes, but the trends are more complicated for Physics 2.

4.3.2.2 How do gender differences in students' high school GPAs, SAT/ACT math scores, low- and high-stakes assessment outcomes differ during remote versus in-person instruction?

There were no statistically significant gender differences in SAT/ACT Math scores in remote or in-person classes for either Physics 1 or Physics 2. Male students tended to have lower high school GPAs than female students, this difference had a medium effect size ($d \sim 0.5$) for both remote and in-person Physics 1 classes (see Table 15). The gender difference in high school GPA had a small ($d \sim 0.2$) effect size for both remote and in-person Physics 2 classes, which can be seen in Table 16. Importantly, sex differences in prior preparation are not drastically different between in-person and remote classes.

Together, homework and quizzes constitute "low-stakes" assessments. Table 15 shows that female students had higher homework scores then male students in Physics 1 for both remote and in-person classes, and the effect sizes of this difference were both small ($d \sim 0.2$) and very similar. On the other hand, Table 16 shows that there was no statistically significant gender difference in homework scores for either in-person or remote courses. There was no statistically significant differences between male and female students' quiz scores in either Physics 1 or Physics 2 in-person or remote classes. For low-stakes assessments, scores either have no statistically significant gender differences, or small gender differences that favor female students.

Midterm and final exams constitute "high-stakes" assessments. Midterm exams did not have any statistically significant gender differences for any class (Physics 1 or Physics 2; in-person or remote), which can be seen in Tables 15 and 16. However, male students tended to have slightly higher midterm exam scores than female students, though the differences were not statistically significant. Final exams had small ($d \sim 0.2$) sex differences for Physics 1 and medium ($d \sim 0.5$) sex differences Physics 2 in-person classes (seen in Tables 15 and 16, respectively). However, remote classes – both Physics 1 and Physics 2 – had no statistically significant sex differences in final exam scores.

4.3.3 Predicting assessment outcomes during remote versus in-person instruction

4.3.3.1 Which factors predict low-stakes assessments scores during remote and in-person instruction?

For this analysis, homework and quizzes were combined into a single low-stakes category. Average homework and quiz scores were weighted equally in this category. Before combining the results, analysis was performed separately for homework and quizzes, which can be found in the supplementary materials.

Regression models predicting low-stakes assessment outcomes in Physics 1 can be found in Table 17. Models 1a and 1b predict low-stakes assessment scores for remote classes, while Models 2a and 2b do the same for in-person courses. We find that average self-efficacy predicts low-stakes assessment outcomes during remote but not in-person classes, which can be seen in Models 1a and 2a in Table 17. Removing average self-efficacy as a predictor, as is done between Models 1a and 1b, lowers the variance in low-stakes assessment outcomes explained. In both remote and in-person courses, low-stakes assessment scores are predicted by high school GPA, but not student sex or test anxiety.

Pre self-efficacy and test anxiety did not predict low-stakes assessment outcomes for Physics 1. For Physics 2, pre self-efficacy and test anxiety either did not predict, or explained very little ($\leq 7\%$) of the variance in low-stakes assessment outcomes. Average self-efficacy and test anxiety did not predict low-stakes grades in Physics 2. All low-stakes regression models not included in Table 17 can be found in the supplementary materials.

4.3.3.2 Which factors predict high-stakes assessment scores during remote and in-person instruction?

For this analysis, midterm and final exams were combined into a single high-stakes category. The instructor gave three midterm exams and one final exam, so the midterm grade made up 75% of the category and the final exam constituted the other 25%. Before combining the results, analysis was performed separately for midterm and final exams, which can be found in the supplementary materials. Models included in the main text use average self-efficacy and test anxiety variables. Models that use only pre self-efficacy and test anxiety predictors can be found in Appendix D, in Tables 63 (Physics 1) and 64 (Physics 2).

Regression models predicting high-stakes assessment outcomes in Physics 1 can be found in Table 18. Models 3a-3d predict high-stakes assessment scores for in-person classes, while Models 4a-4d do the same for remote courses. We find in Model 3a that when both average self-efficacy and test anxiety are predictors of high-stakes assessment outcomes, only self-efficacy, high school GPA, and SAT/ACT math scores predict high-stakes grades. In Model 3a, neither student sex nor test anxiety predict high-stakes grades.

However, Models 3b and 3c show that when included individually, both self-efficacy and test anxiety predict high-stakes assessment scores. This suggests that a mediation model may be appropriate to investigate the relationship between test, anxiety, self-efficacy, and high-stakes assessment scores. Mediation testing can be found in the Figure 5d in Appendix D, and there was a full mediation. That is, though test anxiety predicts high-stakes assessment scores, it's predictive power can be fully directed through self-efficacy.

Finally, in Model 3d, when neither self-efficacy nor test anxiety are included as predictors, student sex becomes statistically significant, which indicates that high-stakes assessment scores only differ by gender when we do not control for self-efficacy or test anxiety; these factors may explain sex differences in high-stakes assessment scores.

Models 4a-4d in Table 18 predict high-stakes assessment scores for Physics 1 remote courses. Similar to in-person courses, Model 4a shows that when both average self-efficacy and test anxiety are predictors of high-stakes assessment outcomes, only self-efficacy, high school GPA, and SAT/ACT math scores predict high-stake grades. In Model 4a, neither student sex nor test anxiety predict high-stakes grades. However, Models 4b and 4c show that when included individually, both self-efficacy and test anxiety predict high-stakes assessment scores. This pattern, similar to the results for in-person courses, suggests that text anxiety may be mediated by self-efficacy. We tested for mediation in Appendix D, and the results can be seen in Figure 5b. Like the mediation models for in-person classes, the relationship between average test anxiety and high-stakes assessment scores is fully mediated by average self-efficacy. For both remote and in-person Physics 1 classes, test anxiety appears to predict self-efficacy, which in turn predicts assessment outcomes.

Next, observing Model 4d, one important difference between remote and in-person classes is that there is no sex difference in high stakes assessment scores, even if we do not control for average self-efficacy or test anxiety.

For Physics 1, the trends are broadly the same for remote and in-person classes. Both self-efficacy and test anxiety predict high-stakes assessment grades, and test anxiety is mediated by self-efficacy when predicting assessment outcomes. However, remote and in-person physics classes differ in some ways. For example, gender predicts high-stakes assessment outcomes when controlling for self-efficacy and test anxiety for in-person classes, but gender does not predict high-stakes assessment during remote classes, regardless of if we controlled for self-efficacy or test anxiety.

Regression models predicting high-stakes assessment outcomes in Physics 2 can be found in Table 19. Models 5a-5d predict high-stakes assessment scores for in-person classes, while Models 6a-6d do the same for remote courses.

We find in Model 5a that when both average self-efficacy and test anxiety are included as predictors of high-stakes assessment outcomes, only self-efficacy, high school GPA, and SAT/ACT math scores statistically significantly predict high-stakes grades. In Model 5a, neither student sex nor test anxiety predict high-stakes grades. However, Models 5b and 5c show that when included individually, both self-efficacy and test anxiety predict high-stakes assessment scores for in-person classes. This pattern again suggests that text anxiety may be mediated by self-efficacy. A mediation model was tested (shown in the Appendix C in Figure 6b), and found that the relationship between average test anxiety and high-stakes assessment scores was fully mediated by average self-efficacy. For both online Physics 1 and 2 classes, test anxiety appears to predict self-efficacy, which in turn predicts assessment outcomes. There are no models in which student sex predicts high-stakes assessment grades for in-person Physics 2 classes.

Models 6a-6c in Table 19 predict high-stakes assessment scores for Physics 2 remote courses. Model 6a shows that when both average self-efficacy and test anxiety are included as predictors of high-stakes assessment outcomes, only self-efficacy, high school GPA, and SAT/ACT math scores statistically significantly predict high-stakes grades. In Model 6a, neither student sex nor test anxiety predict high-stakes grades. Unlike all previous models, test anxiety does not predict high-stakes assessment grades, even if self-efficacy is not included in the model.

Broadly, student sex does not predict high-stakes assessment outcomes for Physics 2 classes. Additionally, the relationship between high school GPA and grades is weaker for Physics 2 than for Physics 1, which can be seen in Models 3a-4d in Table 18 and Models 5a-6c in Table 18. Finally, both self-efficacy and test anxiety predict Physics 2 in-person assessment outcomes, but only self-efficacy predicts these outcomes for remote classes.

4.4 Discussion

4.4.1 Overall differences between in-person versus virtual instruction

Broadly, we found that there were either small ($d \sim 0.2$) or not statistically significant differences in students' test anxiety and self-efficacy between in-person and virtual physics 1 classes. In physics 2, there were small differences in students' pre self-efficacy and post test anxiety: students taking in-person classes started the course with higher self-efficacy and ended the course with more test anxiety than students in virtual classes.

However, the most stark differences between virtual and in-person classes were in assessment outcomes. Students had much higher final exam scores during virtual classes than in-person. The average exam score was 83% in virtual physics 2 class but 57% in the in-person class. Differences were smaller but still significant in midterm exams in both physics 1 and physics 2: students in virtual classes had higher midterm grades than those taking in-person classes. Students taking virtual classes had higher quiz grades in virtual classes for physics 1 but not physics 2. Homework grades were indistinguishable between virtual and in-person classes. Despite minimal differences in self-efficacy and test anxiety, students taking virtual classes had much higher high-stakes assessment grades than their counterparts taking in-person classes. However, students taking virtual and in-person classes had similar low-stakes assessment outcomes.

There are a range of reasons students may have higher high-stakes assessment scores during remote instruction, but we present one hypothesis here. We posit that the exams were lower-stakes during virtual classes because they were in two parts: one group section and one individual section. This may provide an opportunity for students to construct knowledge together during the group portion that they can bring into the individual section [148]. In addition, prior work has found that implementing more frequent exams can improve exam scores [119]. This may be through the mechanism of giving students more frequent access to feedback and to encourage spaced practice (i.e., instead of cramming right before one or two exams, students will study more uniformly), both of which can lead to better retention of content and better skill development [119–122]. Implementing a range of assessment opportunities (such as clicker questions, projects, and group exams) may reproduce an environment in which test anxiety does not correlate as strongly to assessment outcomes, and providing frequent assessments may provide students frequent feedback and opportunities for spaced practice and learning [120, 124].

4.4.2 Gender differences between in-person versus virtual instruction

In all courses, male students tended to report higher levels of self-efficacy and less test anxiety than female students. However, gender differences in both factors tended to decrease from the start to the end of the semester, though students in both groups tended to have lower self-efficacy and more test anxiety at the end of the semester than at the start. There were two exceptions to the trend of sex differences diminishing over time. In-person physics 2 classes had an increased gender difference in test anxiety from the start to the end of the semester, and virtual physics 2 courses had an increased difference in self-efficacy from the start to the end of the semester.

Sex differences in assessment scores tended to be smaller than in motivational factors. During in-person instruction, female students had higher homework scores than male students in physics 1. However, male students had higher midterm exam scores than female students in physics 1 in-person courses, and male students had higher final exam scores than female students in both physics 1 and 2 in-person classes. During virtual instruction, there was small difference in physics 1 homework scores favoring female students, but the average score for male students was also relatively high (88%). There were no other statistically significant grade differences in virtual classes.

Because research suggests that women are more likely than men to leave STEM fields due to concerns about grades (even if they have an A or B average) [115, 156], women's lower exam scores during in-person instruction may contribute to the loss of women from majors that require introductory calculus-based physics. This, combined with recent research that suggests that introductory mathematics courses are better predictors of future course success for physics and engineering students than introductory physics courses [43–45], suggests that many women who may have succeeded in the major leave STEM fields before they have the opportunity to do so.

4.4.3 Predicting assessment outcomes between in-person versus virtual instruction

Generally, test anxiety did not strongly predict low-stakes assessment outcomes, and regression models predicting low-stakes assessment outcomes did not predict much of the variance. High-stakes assessments were predicted by self-efficacy and test anxiety. In both virtual and in-person physics 1 classes, both self-efficacy and test anxiety predict high-stakes assessment grades. However, gender predicted assessment outcomes (even when controlling for self-efficacy, test anxiety, and prior preparation) for in-person classes, but not for virtual classes. In physics 2, both self-efficacy and test anxiety predict physics 2 in-person high-stakes assessment outcomes, but only self-efficacy predicts these outcomes for virtual classes.

For both virtual and in-person physics 1 classes, as well as in-person physics 2 classes, test anxiety is mediated by self-efficacy when predicting high-stakes assessment outcomes. This means that test anxiety may not predict high-stakes assessment outcomes directly, but that test anxiety predicts self-efficacy which in turn predicts high-stakes assessment outcomes. Self-efficacy is built through several mechanisms, but one that may be pertinent here is through management of emotions [63]. Bandura theorized that poor performance and anxiety are 'co-effects' of low self-efficacy [63]. In his paper he theorizes that heightened self-efficacy and reduced test anxiety likely form a virtuous cycle wherein students' development of coping mechanisms (for example, stress-reduction techniques and explicitly rehearsing strategies for academic challenges) increase their self-efficacy, and increased self-efficacy and reduced anxiety frees students' cognitive resources to focus on the task at hand [63, 151].

These models suggest that addressing test anxiety may be a way for instructors to aid their students in building self-efficacy. Thus, creating a low anxiety, equitable, and inclusive learning environment in which all students have a high sense of belonging and feel recognized by their instructors for their effort and progress is important in order for them to master physics and develop confidence in their abilities to do so.

4.5 Conclusion

In this work we investigated differences in motivational factors and learning outcomes between in-person and virtual instruction due to the COVID-19 pandemic. We found that there were small differences in motivational factors, but the differences were most notable in high stakes assessments. The average final exam grade during in-person classes was a 60 in Physics 1 and an 57 in Physics 2. However, the average final exam grade for virtual classes was 78 in Physics 1 and 83 in Physics 2.

Gender differences in motivational factors were diminished more during remote than in-person instruction. Assessment outcomes in high stakes assessments typically favored male students during in-person classes. However, sex differences in exam outcomes were eliminated in high-stakes assessments during virtual classes for both physics 1 and 2.

Finally, we found that both self-efficacy and test anxiety predict high-stakes assessment outcomes in both in-person and virtual classes, though the correlations were weaker for virtual classes. This implies that utilizing methods to increase student self-efficacy and minimize test anxiety may improve student high-stakes grade outcomes.

We acknowledge that one limitation of this study is that it is fundamentally correlational in nature. Any causal relationship between test anxiety, self-efficacy, and learning outcomes would need to be supported, e.g., thaough controlled intervention studies. Another limitation is in the generalizability of our findings. An institution which generally has more low-income students or more students who work full or part-time may have very different virtual course outcomes during COVID than the studied institution. One other limitation of this study is the use of sex rather than gender variables, which may provide a more accurate picture of a student's classroom experiences. We plan to use a gender rather than sex variable as more gender data from this institution becomes available in future years. As we collect more data, we also aim to include intersectional analysis to understand the relationship between gender, race, test anxiety, and self-efficacy.

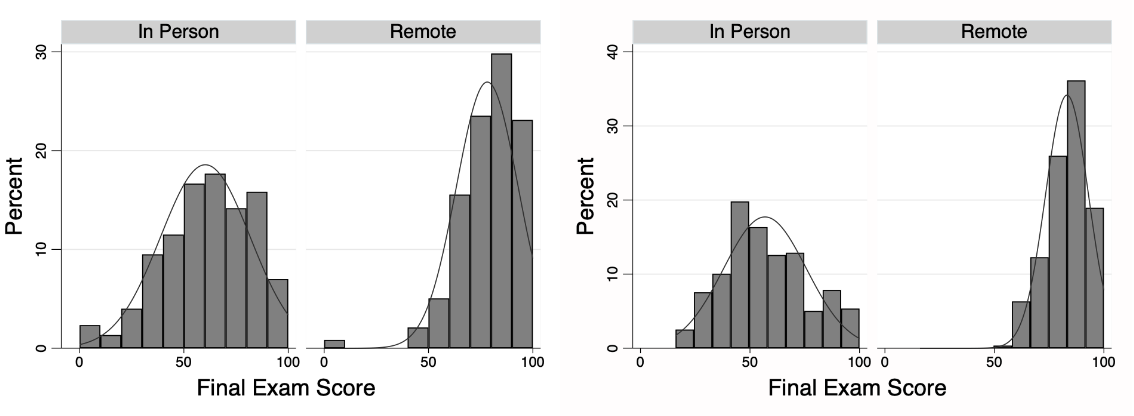
A mediation model for physics 1 virtual classes can be seen in Figures 5c and 5d. Figure 5c shows that test anxiety is statistically significant when predicting high stakes assessment outcomes on its own. However, Figure 5d shows that, as during virtual classes, if test anxiety is used to predict self-efficacy and high-stakes assessment outcomes separately, test anxiety predicts self-efficacy but not high-stakes grades. For virtual physics 1 classes, The average causal mediation effect was 0.33 ($p < 0.001$), with a confidence interval of [0.22, 0.47]. The average direct effect was 0.10, ($p = 0.250$) and the total direct effect was 0.43 ($p < 0.001$).

Table 11: Mean, standard deviation (SD), and comparison of motivational factors, assessment outcomes, and prior preparation of students enrolled in Physics 1 during remote and in-person instruction. Cohen's d effect sizes are also given; a negative d indicates that students taking in-person classes had lower scores then those taking remote classes. $^{ns} = p \geq 0.05$, $^{*} = p < 0.05$, $^{**} = p < 0.01$, and $^{***} = p < 0.001$.

		In-Person			Remote			
Variable	Scale	N	Mean	SD	N	Mean	SD	d
Self-Efficacy Pre	0-1	509	0.68	0.14	218	0.67	0.14	0.10^{ns}
Self-Efficacy Post	0-1	387	0.65	0.18	202	0.66	0.16	-0.06^{ns}
Test Anxiety Pre	0-1	504	0.59	0.26	216	0.56	0.28	0.12^{ns}
Test Anxiety Post	0-1	159	0.54	0.23	204	0.48	0.29	0.22^{*}
Homework	0-100	600	90	16	238	90	18	0.00^{ns}
Quizzes	0-100	600	87	12	238	92	10	-0.46^{***}
Midterm Exams	0-100	600	74	16	238	78	13	-0.31^{***}
Final Exam	0-100	600	60	22	238	78	15	-0.90^{***}
SAT/ACT Math	200-800	503	706	61	238	706	59	0.00^{ns}
High School GPA	0-5	600	4.19	0.37	238	4.25	0.36	-0.18^{*}

Table 12: Mean, standard deviation (SD), and comparison of motivational factors, assessment outcomes, and prior preparation of students enrolled in Physics 2 during remote and in-person instruction. Cohen's d effect sizes are also given; a negative d indicates that students taking in-person classes had lower scores then those taking remote classes. $^{ns} = p \geq 0.05$, $^{*} = p < 0.05$, $^{**} = p < 0.01$, and $^{***} = p < 0.001$.

		In-Person			Remote			
Variable	Scale	N	Mean	SD	N	Mean	SD	d
Self-Efficacy Pre	0-1	206	0.72	0.15	253	0.67	0.17	0.33^{***}
Self-Efficacy Post	0-1	251	0.70	0.16	130	0.67	0.15	0.20^{ns}
Test Anxiety Pre	0-1	209	0.51	0.25	252	0.53	0.28	-0.08^{ns}
Test Anxiety Post	0-1	258	0.47	0.25	129	0.53	0.26	-0.24^{*}
Homework	0-100	318	94	11	285	94	13	-0.02^{ns}
Quizzes	0-100	318	91	10	285	90	8	0.13^{ns}
Midterm Exams	0-100	318	77	10	285	85	8	-0.89^{***}
Final Exam	0-100	318	57	19	285	83	10	-1.71^{***}
SAT/ACT Math	200-800	261	716	59	285	720	52	-0.07^{ns}
High School GPA	0-5	318	4.31	0.34	285	4.34	0.30	-0.09^{ns}



(a) Physics 1

(b) Physics 2

Figure 4: Histograms of student final exam grades with an overlaid normal curve. Figure (a) shows the grades for Physics 1 for online and in-person classes, while (b) does the same for Physics 2.

Table 13: Self-efficacy and test anxiety survey scores for male and female students taking both remote and in-person Physics 1 courses. Cohen's d effect sizes are also given; a negative d indicates that men had lower scores then women. $^{ns} = p \geq 0.05$, $^{*} = p < 0.05$, $^{**} = p < 0.01$, and $^{***} = p < 0.001$.

			Female			Male			
	Variable	Pre/Post	N	Mean	SD	N	Mean	SD	d
In-Person	Self-Efficacy	Pre	179	0.63	0.13	330	0.71	0.13	0.55^{***}
		Post	142	0.60	0.18	245	0.68	0.17	0.42^{***}
	Test Anxiety	Pre	177	0.47	0.27	327	0.66	0.23	0.75^{***}
		Post	55	0.48	0.23	104	0.57	0.23	0.41^{*}
Remote	Self-Efficacy	Pre	92	0.62	0.15	126	0.70	0.11	0.59^{***}
		Post	92	0.64	0.15	110	0.68	0.15	0.27^{ns}
	Test Anxiety	Pre	92	0.46	0.29	110	0.63	0.25	0.63^{***}
		Post	92	0.42	0.29	110	0.53	0.29	0.38^{**}

Table 14: Self-efficacy and test anxiety survey scores for male and female students taking both remote and in-person Physics 2 courses. Cohen's d effect sizes are also given; a negative d indicates that men had lower scores then women. $^{ns} = p \geq 0.05$, $^{*} = p < 0.05$, $^{**} = p < 0.01$, and $^{***} = p < 0.001$.

			Female			Male			
	Variable	Pre/Post	N	Mean	SD	N	Mean	SD	d
In-Person	Self-Efficacy	Pre	76	0.68	0.13	130	0.74	0.15	0.37^{**}
		Post	84	0.67	0.16	167	0.71	0.15	0.27^{*}
	Test Anxiety	Pre	77	0.45	0.23	130	0.54	0.25	0.36^{*}
		Post	86	0.38	0.21	172	0.51	0.26	0.55^{***}
Remote	Self-Efficacy	Pre	109	0.63	0.16	144	0.69	0.16	0.37^{**}
		Post	53	0.61	0.13	77	0.70	0.15	0.62^{***}
	Test Anxiety	Pre	108	0.44	0.25	144	0.60	0.28	0.59^{***}
		Post	52	0.44	0.26	77	0.59	0.25	0.59^{**}

Table 15: Mean and standard deviartion (SD) of assessment and prior preparation scores for male and female students taking both remote and in-person Physics 1 courses. Cohen's d effect sizes are also given; a negative d indicates that men had lower scores then women. $^{ns} = p \geq 0.05$, $^{*} = p < 0.05$, $^{**} = p < 0.01$, and $^{***} = p < 0.001$.

			Female			Male			
	Variable	Scale	N	Mean	SD	N	Mean	SD	d
In-Person	SAT/ACT Math	200-800	164	702	65	339	708	59	0.10^{ns}
	HS GPA	0-5	203	4.29	0.34	397	4.13	0.38	-0.45***
	Homework	0-100	203	93	13	397	89	17	-0.27**
	Quizzes	0-100	203	87	12	397	87	12	-0.03^{ns}
	Midterm Exams	0-100	203	71	15	397	75	16	0.25**
	Final Exam	0-100	203	56	21	397	62	22	0.29***
Remote	SAT/ACT Math	200-800	101	699	54	137	712	54	0.24^{ns}
	HS GPA	0-5	101	4.37	0.31	137	4.17	0.37	-0.57***
	Homework	0-100	101	93	14	137	88	20	-0.30*
	Quizzes	0-100	101	93	7	137	91	11	-0.22^{ns}
	Midterm Exams	0-100	101	79	11	137	77	14	-0.15^{ns}
	Final Exam	0-100	101	77	14	137	79	16	0.09^{ns}

Table 16: Mean and standard deviartion (SD) of assessment and prior preparation scores for male and female students taking both remote and in-person Physics 2 courses. Cohen's d effect sizes are also given; a negative d indicates that men had lower scores then women. $^{ns} = p \geq 0.05$, $^{*} = p < 0.05$, $^{**} = p < 0.01$, and $^{***} = p < 0.001$.

			Female			Male			
	Variable	Scale	N	Mean	SD	N	Mean	SD	d
In-Person	SAT/ACT Math	200-800	76	720	55	185	714	60	-0.10^{ns}
	HS GPA	0-5	101	4.39	0.30	217	4.27	0.34	-0.35**
	Homework	0-100	101	96	9	217	93	12	-0.27^{ns}
	Quizzes	0-100	101	92	8	217	91	10	-0.15^{ns}
	Midterm Exams	0-100	101	75	11	217	78	10	0.22^{ns}
	Final Exam	0-100	101	52	19	217	59	18	0.40***
Remote	SAT/ACT Math	200-800	124	720	52	161	720	53	-0.02^{ns}
	HS GPA	0-5	124	4.39	0.28	161	4.30	0.30	-0.30*
	Homework	0-100	124	96	11	161	93	14	-0.17^{ns}
	Quizzes	0-100	124	90	7	161	90	9	-0.03^{ns}
	Midterm Exams	0-100	124	85	9	161	86	8	0.11^{ns}
	Final Exam	0-100	124	83	9	161	84	10	0.07^{ns}

Table 17: Physics 1 low-stakes assessment scores predicted by student sex, High School GPA (HS GPA), SAT/ACT Math scores, average self-efficacy and average test anxiety. Standardized regression (β) coefficients are provided. $^{*} = p < 0.05$, $^{**} = p < 0.01$, $^{***} = p < 0.001$, and ^{ns} = not statistically significant.

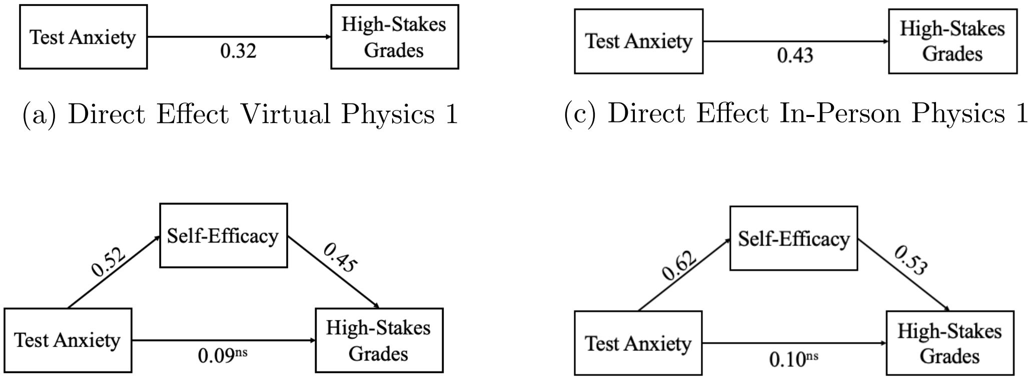
	Remote (N=187)		In-Person (N=146)	
Variable	Model 1a	Model 1b	Model 2a	Model 2b
Sex	0.11^{ns}	0.06^{ns}	0.11^{ns}	0.07^{ns}
HS GPA	0.25^{**}	0.27^{***}	0.29^{***}	0.31^{***}
SAT/ACT Math	0.08^{ns}	0.14^{*}	0.07^{ns}	0.09^{ns}
Self-Efficacy	0.27^{***}		0.15^{ns}	
Test Anxiety	-0.04^{ns}		-0.05^{ns}	
Adjusted R^2	0.15	0.10	0.11	0.11

Table 18: Physics 1 high-stakes assessment scores predicted by student sex, High School (HS) GPA, SAT/ACT Math scores, average self-efficacy, and average test anxiety. Adjusted R^2 and standardized regression (β) coefficients are provided. For remote classes, $N = 187$, and for in-person classes, $N = 146$. $^{ns} = p \geq 0.05$, $^{a} = p < 0.05$, $^{b} = p < 0.01$, and $^{c} = p < 0.001$.

	Model	Sex	HS GPA	SAT/ACT	SE	TA	Adj. R^2
In-Person	Model 3a	0.00^{ns}	0.28^{c}	0.35^{c}	0.45^{c}	0.11^{ns}	0.55
	Model 3b	-0.02^{ns}	0.27^{c}	0.36^{c}	0.50^{c}		0.55
	Model 3c	-0.06^{ns}	0.34^{c}	0.38^{c}		0.36^{c}	0.43
	Model 3d	-0.18^{b}	0.32^{c}	0.46^{c}			0.33
Remote	Model 4a	0.10^{ns}	0.21^{b}	0.26^{c}	0.37^{c}	0.10^{ns}	0.38
	Model 4b	0.08^{ns}	0.21^{b}	0.26^{c}	0.42^{c}		0.37
	Model 4c	0.07^{ns}	0.22^{b}	0.33^{c}		0.27^{c}	0.28
	Model 4d	-0.02^{ns}	0.24^{b}	0.37^{c}			0.22

Table 19: Physics 2 high-stakes assessment scores predicted by student sex, High School (HS) GPA, SAT/ACT Math scores, as well as average self-efficacy and test anxiety. Adjusted R^2 and standardized regression (β) coefficients are provided. For remote classes, $N = 187$ and for in-person classes, $N = 146$. $^{ns} = p \geq 0.05$, $^{a} = p < 0.05$, $^{b} = p < 0.01$, and $^{c} = p < 0.001$.

	Model	Sex	HS GPA	SAT/ACT	SE	TA	Adj. R^2
In-Person	Model 5a	-0.02^{ns}	0.14^{a}	0.38^{c}	0.39^{c}	0.00^{ns}	0.40
	Model 5b	-0.04^{ns}	0.13^{ns}	0.44^{c}	0.39^{c}		0.40
	Model 5c	-0.02^{ns}	0.14^{a}	0.38^{c}		0.18^{a}	0.29
	Model 5d	-0.07^{ns}	0.11^{ns}	0.49^{c}			0.26
Remote	Model 6a	0.05^{ns}	0.12^{ns}	0.32^{c}	0.27^{b}	0.05^{ns}	0.22
	Model 6b	0.05^{ns}	0.13^{ns}	0.32^{c}	0.29^{b}		0.23
	Model 6c	-0.06^{ns}	0.17^{ns}	0.37^{c}			0.16



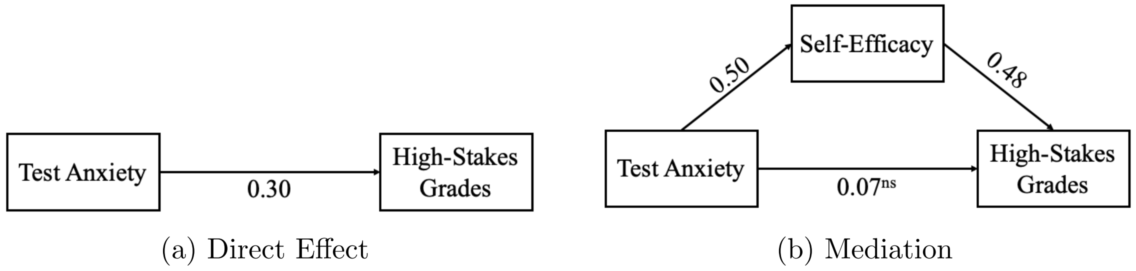
(a) Direct Effect Virtual Physics 1

(c) Direct Effect In-Person Physics 1

(b) Mediation Virtual Physics 1

(d) Mediation In-Person Physics 1

Figure 5: Mediation model results for physics 1 classes. For virtual classes ($N = 187$), (a) shows the model without self-efficacy, while (b) shows the model including self-efficacy. For in-person classes ($N = 147$), (c) shows the model without self-efficacy, while (d) shows the model including self-efficacy. Average self-efficacy mediates the relationship between average test anxiety and high-stakes assessment scores. Unless specified, all regression coefficients are significant to the $p < 0.001$ level. ^{ns} indicates a result is not statistically significant.



(a) Direct Effect

(b) Mediation

Figure 6: Mediation model results for physics 2 in-person classes ($N = 180$): (a) shows the model without self-efficacy, while (b) shows the model including self-efficacy. Average self-efficacy mediates the relationship between average test anxiety and high-stakes assessment scores. Unless specified, all regression coefficients are significant to the $p < 0.001$ level. ^{ns} indicates a result is not statistically significant.

5.0 Peer interaction, self-efficacy, and equity: Same gender groups are more beneficial than mixed gender groups for female students

5.1 Introduction and Theoretical Framework

In recent years, science education researchers have particularly focused on strategies for creating equitable and inclusive learning environments for underrepresented students such as women [39, 44, 157, 158]. Here, we focus on women because they are drastically underrepresented in science disciplines that require students to take calculus-based physics, such as physics, chemistry and engineering [1, 159]. Many frameworks have been proposed for understanding gender disparities in science, technology, engineering, and mathematics (STEM) courses. For example, motivational factors such as self-efficacy are often used to investigate gender differences in physics performance and persistence [10, 160, 161].

According to Bandura, self-efficacy is one's belief in their capability to succeed at an activity or subject [127], and high self-efficacy correlates with positive grade and retention outcomes for students [8, 12, 32]. Self-efficacy is developed through four mechanisms [64, 127]. The first is mastery experiences, which describes learning by overcoming difficulties such as a challenging problem set. The second is social modeling [127]. This describes seeing people similar to oneself succeeding in a field of study (for example, somebody who shares your gender). The third is social persuasion, which is encouragement to increase resolve and measure success via personal improvement [127]. The final mechanism is emotional state, such as management of anxiety [127]. Partly due to pervasive stereotypes and biases pertaining to who belongs in physics and can excel in it, women taking physics courses often have lower physics self-efficacy than men [11, 23, 41, 44, 65]. Another reason for this gender disparity may be that women tend to have fewer social modeling experiences in the context of physics because women in physics have fewer peers and role models (e.g., instructors or teaching assistants and well-known researchers) that share their gender [1, 162, 163].

Prior research suggests that group work, even if all members have similar levels of knowledge, can positively affect individual learning outcomes [148, 164, 165]. Although most of the prior research focuses on how working with peers can directly improve student learning, peers can also affect students' motivational beliefs [98, 99, 166]. This is particularly important because of the long-term benefits of increased self-efficacy [167–169]. For example, women in engineering courses experience less anxiety and participate more if they work in majority-female groups compared to majority-male groups [98]. Same-gender peer mentoring also has a positive impact on the self-efficacy of women in engineering programs [99,162]. We posit that the women in these studies have an increase in self-efficacy due to the social modeling experiences facilitated by working with other women and reduced stereotype threat (stereotype threat can increase if students from the dominant group, e.g., men in physics, dominate the discussions in a mixed-gender group work). In the physics context, one study found that, in a group of two men and one woman, the woman's inputs tended to be disregarded even if she was deemed by the researcher to have the highest level of physics knowledge in the group [170]. Additionally, women may be disadvantaged in mixed-gender groups compared to men due to gendered task division [34, 35]. There is also support from individual interviews that women feel more confident working with other students of the same gender, while men claim they are comfortable working with either gender [171].

In addition to measuring self-efficacy directly, here we introduce a new measure:

self-reported “peer influence on self-efficacy”, which measures students’ perceptions about how interactions with their peers affected their confidence in physics. This measure may be especially important for high-enrollment classes in which students interact much more with other students than with the instructor or teaching assistant (e.g., when they participate in clicker questions with peer discussion or solve problems collaboratively in recitation classes). We hypothesize that self-reported peer influence on self-efficacy may also be a useful measure of how equitable and inclusive those peer interactions are. In this study, we aim to answer the following research questions:

RQ1. How do women and men’s self-efficacy and peer influence on self-efficacy compare within the three subgroups (students who worked alone, students who worked in same-gender groups, and students who worked in mixed-gender groups)?

RQ2. How does self-efficacy change over time for students in each subgroup?

RQ3. ow does gender and group type predict student self-efficacy and peer influence on self-efficacy?

The answers to these questions have important implications on how instructors organize group work and create an equitable learning environment in which group work is valuable for all students regardless of their gender.

5.2 Methods

5.2.1 Participants and Procedures

Participants were students who enrolled in the first of the two calculus-based introductory physics course during two consecutive fall semesters. This traditional lecture-based course covers Newtonian mechanics. Most students enrolled in the course are first-year engineering or physical science (chemistry and physics) majors. Surveys were completed in recitation during the first and last week of the semester, which we call “pre” and “post” surveys, respectively. Students were offered either course credit or extra credit for completion, depending on the instructor’s preference. There were 1266 students who completed the pre-survey and 930 who completed the post-survey. The discrepancy is partially due to some (N=148) students being given a different post survey. Also, students may miss the first or last recitation of the semester when the survey was administered for many reasons. Less than 1% did not list their gender, and less than 1% did not pass the survey’s attention check (a random question number that requested students respond with “C”). Three groups were were excluded from the study, leaving 890 total students in our study who completed both surveys. The sample was 37% women and 63% men. This research was carried out in accordance with the principles outlined in the University of Pittsburgh Institutional Review Board (IRB) ethical policy, and de-identified demographic data were provided through university records.

5.2.2 Measures and Survey Validation

The survey was designed to measure students’ physics self-efficacy and peer influence on self-efficacy, and was adapted from previously validated surveys [9, 41]. We conducted 20 one-hour semi-structured interviews with current and former physics students to ensure that students interpreted questions as intended. A few questions were edited slightly after the interviews during the validation process, and the questions in the final set were interpreted by students as intended.

Self-efficacy is one's belief in their capability to succeed at an activity or subject [127]. Students' self-reported peer influence on self-efficacy (PISE) is how students believe their peer interactions influenced their self-efficacy. In our survey, we included four physics self-efficacy and four PISE items, each on a four-point Likert scale (1—NO!, 2—no, 3—yes, 4—YES!). These items are listed in Table 20.

After initial exploratory factor analysis that included other motivational constructs not discussed here that were part of the same survey instrument, we performed Confirmatory Factor Analysis (CFA) for each construct. We considered the model a good fit if it met certain cutoffs for various fit indices. These cutoffs are: comparative fit index (CFI) ≥ 0.9 and Tucker Lewis index (TLI) ≥ 0.9 [172]. Additionally, root mean square error of approximation (RMSEA) ≤ 0.08, and standardized root mean square residual (SRMR) ≤ 0.08 are considered an acceptable fit [172]. We also conducted measurement invariance tests to determine if men and women could be included in the same models, and found that they could. The results of the CFA can be seen in Table 20. The factor loadings for all constructs were all greater than the out cutoff of 0.5 [173], and Cronbach's α for each construct was between 0.7 and 0.95 [108]. Finally, students were asked "Most typically in this physics course ..." with three answer options: "I worked alone", "I worked with students mostly of my own gender", and "I worked with students mostly of another gender." Because each instructor structured their course differently, there was no standard set of experiences students had within their groups. Instead, the question is likely to reflect their experiences working on some combination of clicker questions, in-class group work, homework, study groups, and office hours, depending on the structure of the course. The number of students of each gender that worked in each sort of group can be seen in Table 21.

Table 20: Survey items for each of the motivational scales. The model fit indices were as follows: CFI = 0.98, TLI = 0.97, RMSEA = 0.067, and SRMR = 0.026. The factor loadings (λ) are all standardized and significant to the $p < 0.001$ level.

Construct and Item	λ
Self-Efficacy, $\alpha = 0.81$	
I am able to help my classmates with physics in the laboratory or in recitation.	0.73
I understand concepts I have studied in physics.	0.71
If I study, I will do well on a physics test.	0.75
If I encounter a setback in a physics exam, I can overcome it	0.67
Peer Influence on Self-Efficacy, $\alpha = 0.92$	
My experiences and interactions with other students in this class...	
made me feel more relaxed about learning physics.	0.73
increased my confidence in my ability to do physics.	0.90
increased my confidence that I can succeed in physics.	0.93
increased my confidence in my ability to handle difficult physics problems.	0.88

5.2.3 Analysis

First, we calculated the mean scores for pre self-efficacy, post self-efficacy, and peer influence on self-efficacy. Peer influence on self-efficacy was only measured at the end of the semester on the post survey. We found these values for men and women separately within each of three student categories: those who worked alone, those who worked in same gender groups (SGGs), and those who worked in mixed gender groups (MGGs). Then we used a regular two-sample t-test [108] to compare responses of female and male students.

We used structural equation modeling (SEM) to study the effect of student gender and pre self-efficacy on post self-efficacy and peer influence on self-efficacy. We performed a multigroup SEM to explore the model differences for each group of students. By performing a full SEM, we combine CFA with path analysis, which provides regression coefficients between multiple factors while allowing for multiple outcomes. We use the same fit indices and cutoffs for full SEM as we do for CFA. All analysis was conducted using R [174] and the package lavaan [175].

Table 21: Number and percentage of students working alone, in same gender groups, and in mixed gender groups.

All Students	Alone	Same Gender Group	Mixed Gender Group
Women	N=74 (23%)	N=58 (18%)	N=194 (60%)
Men	N=183 (33%)	N=109 (20%)	N=258 (47%)

5.3 Results and Discussion

With regard to RQ1, regardless of group type, men in our sample have higher pre self-efficacy, post self-efficacy (as seen in Table 22), and peer influence on self-efficacy than women (as seen in Table 23). In both Tables 22 and 23, a high score (i.e., close to 4) for self-efficacy or peer influence on self-efficacy indicates that the student has high self-efficacy or a more positive peer influence on self-efficacy. This result supports past research that shows that men often have higher self-efficacy than women, especially in physics [41, 113, 176]. One major reason for this self-efficacy difference is societal stereotypes and biases about who belongs in physics and can excel in physics and other STEM fields [58, 59, 177].

We find, using Table 22, that the largest gender difference in pre self-efficacy is between students who work in MGGs (Cohen's d=0.57, $p < 0.041$). This is because men who worked in MGGs tended to have higher pre self-efficacy than men who work alone or in SGGs, while women who work in MGGs do not have higher pre self-efficacy than women who work alone or in SGGs. The largest gender difference in post self-efficacy is also between students who worked in MGGs. One hypothesis for why this may be the case is that group work between men with higher self-efficacy and women with lower self-efficacy created inequitable and non-inclusive learning environments in which men dominated the group work [38, 178–180].

There was also a gender difference in reported peer influence on self-efficacy, seen in Table 23, regardless of group type. One potential reason for this difference is that students are influenced by societal stereotypes and biases when they enter the physics classroom: societal stereotypes about women's abilities in the sciences affect not only how they are treated by parents [97, 181] and instructors [56], but also how they are treated by peers [182]. Examples of interactions that may decrease women's reported peer influence on self-efficacy include having ideas ignored by men in their classes [50], inequitable group roles such as being expected to write down answers or manage time instead of focusing on physics concepts [34], and experiencing sexist comments from peers [183].

In response to RQ2, we explore how self-efficacy may change over time for students in each subgroup. Concerningly, students who worked alone or in MGGs tended to have significant decreases in self-efficacy over the semester, and those decreases had larger effect sizes for women than men. In particular, we find that men who work alone or in MGGs have a statistically significant but marginal (Cohen's $|d| < 0.1$)

drops in self efficacy, but men who work in same gender group have no significant decrease in self-efficacy from pre to post, as seen in Table 22. We also find that women have significant but small (Cohen's $|d|$ ~0.2) decreases in self efficacy if they work alone or in MGGs. However, women have no significant self-efficacy decrease from pre to post if they work in SGGs.

Thus, our findings suggest that the classroom environment generally decreased students' self-efficacy. However, students who worked in SGGs did not have significant drops in self-efficacy from the beginning to end of the semester. Therefore, the dynamics of same gender group work may be a good example for instructors to take inspiration from in order to create equitable and inclusive learning environments and minimize self-efficacy decreases which can be particularly detrimental for the underrepresented students such as women.

One area that needs further investigation is the relationship between self-efficacy and peer influence on self-efficacy. Men who worked in MGGs had a small, borderline-significant decrease in self-efficacy from pre to post, and they reported the highest peer influence on self-efficacy. This result is intuitive: if students' peers have a positive effect on their self-efficacy, then their self-efficacy should drop less. However, women who worked in MGGs had a slightly larger decrease in self-efficacy compared to those who worked alone or in SGGs, but reported higher peer influence on self-efficacy than other women. While we cannot know the reason for this without interviewing these women, we have some potential hypotheses for why this may be the case as follows.

If women who work in MGGs work with men who appear to be friendly, then during group work they may feel supported by their peers, even if their group work practices have inequitable outcomes [36]. In this sort of group, women will have mastery experiences as they complete assignments, and they may also have peers who provide social persuasion that increases self-efficacy. These are two modes of increasing self-efficacy [127]. However, due to gendered task division [34, 35] and women's assumed lack of ability by men [183], women in mixed gender groups may be less likely to witness social modeling from other women. Another method of increasing self-efficacy is the management of emotions, such as anxiety [127]. Women are more likely to experience anxiety in academic settings [60] due to societal biases against women in STEM fields [59] and stereotype threat [116] which may be triggered by mixed-gender group work particularly if men dominate. Thus, women working in MGGs may have fewer opportunities to develop self-efficacy than those who work in SGGs, even if they have supportive group members. We emphasize here that an equitable learning environment is not only an environment in which minoritized students do not experience blatant discrimination, but one in which all students achieve course objectives and personal goals, regardless of background. Because self-efficacy is correlated with increased learning outcomes [184], enrollment in future physics classes [51], and career choices [12], implementing classroom policies that eliminate this gendered self-efficacy difference is an important part of creating an equitable and inclusive environment in physics classes and departments.

Finally, in response to RQ3, we explore how gender and group type predict student self-efficacy and peer influence on self-efficacy. We conducted multigroup SEM, seen in Figure 1, to produce three similar path analysis models, one each for students who worked alone, students who worked in same gender groups, and students who worked in mixed gender groups. Each model shows the predictive relationship of gender on post self-efficacy and peer influence on self-efficacy, allowing us to compare this relationship across group types. The model fit indices suggest a good fit to the data: CFI = 0.98 (≥ 0.95), TLI = 0.98 (≥ 0.95), RMSEA = 0.057 (≤ 0.06), and SRMR = 0.029 (≤ 0.06).

In Figure 7, the solid lines are labeled with standardized regression coefficients (β). Standardized regression coefficients represent the expected change in an outcome variable for each change in standard deviation of an input variable, while controlling for other variables in the model. For example, if β=0.62 between peer effect on self-efficacy and post self-efficacy, then if a student' peer effect on self-efficacy score is one standard deviation higher than the mean, we would expect their post self-efficacy score to be 0.62 standard deviations higher than the mean. From these models, we note several important findings. First, for students who work alone, gender predicts neither their post self-efficacy nor peer influence on self-efficacy.

Though gender differences are smallest in this group, this is likely because women who work alone start with slightly higher self-efficacy than other women (see Table 22), rather than due to classroom practices. Women who work alone have similar self-efficacy decreases to women who work in MGGs, and larger self-efficacy decreases than women who work in SGGs.

We note that gender predicts peer influence on self-efficacy, for both SGGs and MGGs. This suggests that men's self-efficacy tends to benefit more from group work than women's. Additionally, the only group for which gender significantly predicts post self-efficacy is students who worked in MGGs. This suggests that women's post self-efficacy may be harmed by mixed gender group work in inequitable learning environments. This may be due to aforementioned inequitable group work practices.

5.3.1 Implications

Based on the results of this research, we suggest that same-gender group work is especially beneficial to women in these classes. This is particularly important for women, who may feel more confident working with other students of the same gender [171]. Prior work suggests that having the women-only groups increases engineering students' self-efficacy [99]. One explanation may be that working in majority-female groups may provide students with "stereotype inoculation", which may directly combat internalized stereotypes about women's ability in physics by providing positive female role models, providing evidence to students with low self-efficacy that their gender will not prevent them from succeeding [98, 99, 162].

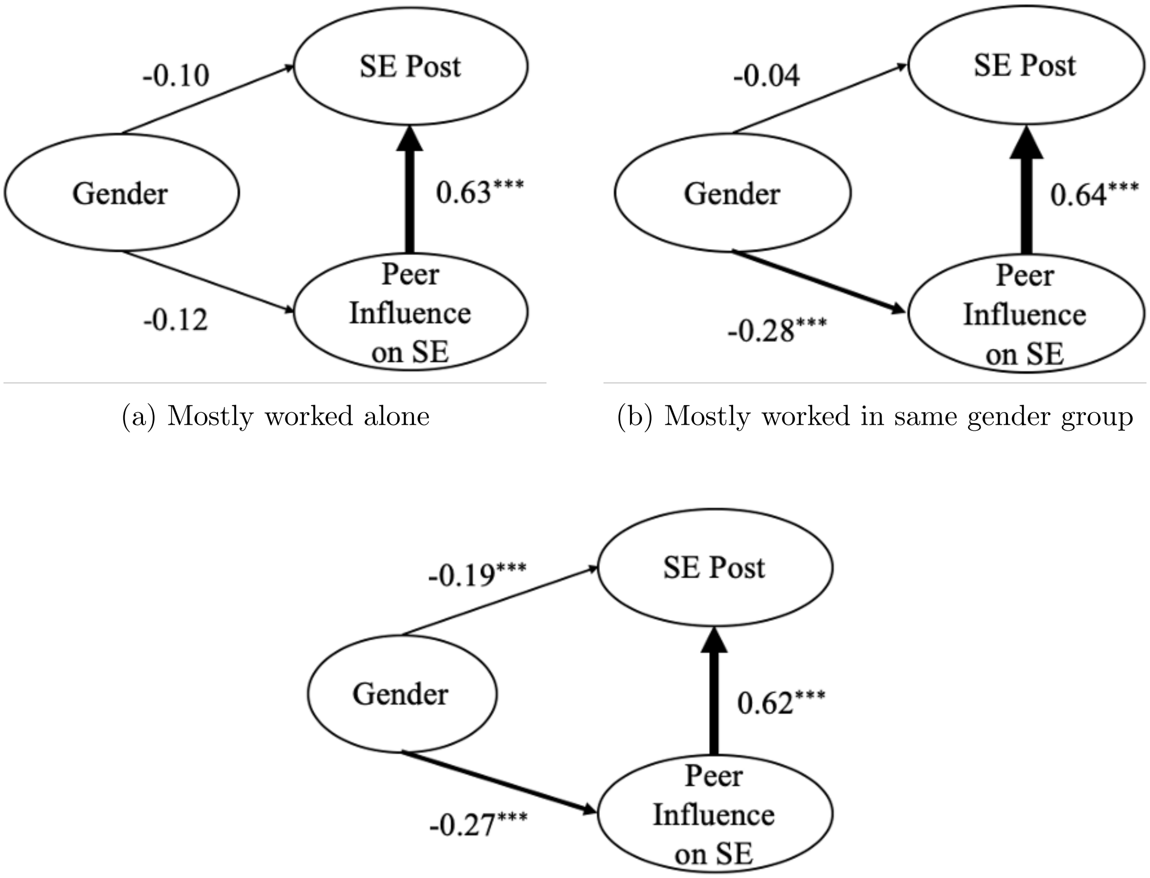
However, instructors should not discourage students from working in MGGs, especially if students' alternative is working alone. This is due to the body of evidence that group work is beneficial to students and students can even co-construct knowledge when each student did not know how to solve a problem but students were able to figure it out together [148, 164, 165]. Instead, instructors should implement policies that encourage a more equitable and inclusive group environment. For example, implementing individual accountability for students working in groups (e.g., reviewing peer interactions based on inclusive teamwork practices and task division) can be valuable to avoid an inequitable dynamic in which a woman in a group acts as a "secretary" who takes notes or becomes a "manager" for the men she is working with (being a manager of the group can take up cognitive resources she could be using to learn since working memory is limited) [34]. Another possibility to increase gender equity in group work is creating opportunities for all students to act as "experts" in a subject, for example assigning students to teach a topic to a small group of peers. Assuming the role of an expert can increase students' self-efficacy [185], and this may be especially beneficial for women since it can give all of the women in the classroom an opportunity to build self-efficacy through social modeling.

5.3.2 Limitations and Future Directions

This research is based at a primarily white, large, public university. While our results may generalize to similar institutions, it is important to see if similar patterns exist at smaller liberal arts colleges, minority-serving institutions, or community colleges in the US. Additionally, though we investigate group work more generally here, future research can focus on specific group work situations (for example, structured in-class group work or informal study groups). Finally, future investigation can study if reverse-coding some items may change the results of the study in order to further improve the PISE measure.

5.4 Conclusion

In this study, we found that men tended to have higher self-efficacy and reported higher peer influence on self-efficacy than women, regardless of whether they worked alone, in same gender groups, or in mixed gender groups. This disparity may be due to societal stereotypes and biases about who can excel in science, particularly in physics. Further, we found that students who worked in same-gender groups had a non-significant change in self-efficacy from the beginning to the end of the semester as measured by the pre to post survey. However, students who worked alone or in mixed gender groups had significant drops in self-efficacy. Furthermore, SEM shows that gender predicts both peer effect on self-efficacy and self-efficacy if students worked in mixed gender groups, but gender does not predict self-efficacy if students worked in a same gender group. Instructors can take inspiration from the group practices in the same gender group to increase equity and inclusion in group work overall. In particular, instructors should implement classroom policies that encourage equitable and inclusive group work, so that all students can benefit regardless of their demographic group.



(a) Mostly worked alone

(b) Mostly worked in same gender group

(c) Mostly worked in mixed gender group

Figure 7: Multigroup path analysis models for students who worked alone, in same gender outcomes groups, and in mixed gender groups, predicting the effect of gender and peer influence on self efficacy (SE) on self-efficacy. The lines represent regression paths, and the line thickness corresponds to the magnitude of β value (standardized regression coefficient) with $p < 0.001$ indicated by ***. If a coefficient is nonsignificant, it has no asterisks.

Table 22: Mean scores of pre and post self-efficacy. Cohen's d is used to compare the effect size between men and women for each group (worked alone, worked in a same-gender group, worked in a mixed-gender group) as well as to compare pre and post self-efficacy for men and women in each group. A positive d indicates that men had higher scores than women or that self-efficacy increased from the beginning to end of the semester. The p-value reports the significance level of the t-tests comparing either men and women or the pre and post results of each gender group.

		SE Pre	SE Post	Statistics	
	Gender	1-4	1-4	Cohen's *d*	*p*-value
Alone	Women (N=74)	2.89	2.78	-0.21	0.021
	Men (N=183)	3.01	2.97	-0.08	0.011
	Cohen's *d*	0.24	0.36		
	p-value	0.080	0.010		
Same Gender Group	Women (N=58)	2.78	2.72	-0.10	0.263
	Men (N=109)	2.99	2.94	-0.11	0.138
	Cohen's *d*	0.46	0.43		
	p-value	0.005	0.010		
Mixed Gender Group	Women (N=194)	2.84	2.72	-0.24	¡0.001
	Men (N=258)	3.12	3.08	-0.08	0.050
	Cohen's *d*	0.57	0.70		
	p-value	0.041	¡0.001		

Table 23: Mean scores of peer influence on self-efficacy (PISE). Cohen's d is used to compare the effect size between men and women for each group (worked alone, worked in a same-gender group, worked in a mixed-gender group). A positive d indicates that men had higher scores than women. The p-value reports the significance level of the between-gender t-tests.

	Women		Men		Statistics	
	N	PISE (1-4)	N	PISE (1-4)	Cohen's d	p-value
Alone	74	2.64	183	2.86	0.31	0.027
Same Gender Group	58	2.65	109	3.01	0.28	<0.001
Mixed Gender Group	194	2.74	258	3.04	0.44	<0.001

6.0 Whose ability and growth matter? Gender, mindset and performance in physics

6.1 Introduction

Improving the diversity in post-secondary science, technology, engineering, and mathematics (STEM) education has been a long-standing focus of policymakers and researchers [4, 186] Physics and engineering in particular have very low numbers of women in high school courses, undergraduate programs, and in the fields [1, 66].

Numerous factors affect representation in STEM fields. For example, parents of girls are less likely to believe their child could succeed in a career that requires mathematical ability [67, 97]. Once they are in high school, girls are less likely than boys to believe that a career in physics could align with their professional goals [185]. In physics, there are also gender disparities in introductory course performance [81–83, 187]. Motivational factors have been linked to general academic performance [6–8] as well as to gender differences in persistence in STEM courses [12, 115, 188] and performance in STEM courses [11, 22, 23, 40, 41, 65]

Among the many motivational factors that have been investigated, researchers have put considerable attention on the role of intelligence mindsets [128, 189] Intelligence mindset describes a person's beliefs about the nature of intelligence: is it innate and unchangeable or something that can be developed with effort [68]? In more recent years, a focus has shifted to discipline-specific intelligence mindsets since students appeared to have separate views by discipline and the discipline-specific mindset was more predictive of student performance in the discipline [33, 40]. However, since physics-specific mindset is still a very recently explored concept, many fundamental questions about its nature and relationship to gendered performance in physics are still open. Specifically, we address the research questions:

RQ 1. What are the components to students' physics intelligence mindsets?

RQ 2. Are there gender/sex differences in the components of students' physics intelligence mindsets?

RQ 3. If there are differences in the components of students' physics intelligence mindsets, do the differences grow or decline from the beginning to the end of their first university-level physics course?

RQ 4. Do any of the mindset components predict course grade?

To answer these questions, we chose the first calculus-based introductory course as the research context. Introductory calculus-based physics courses are typically taken by engineering and physical science majors, while most algebra-based physics students are life science and pre-medical majors. As a result, calculus-based introductory physics courses are likely to be majority men, which likely further reinforces stereotypes and negative messages that women in physics courses are receiving [1,66,190]. Because of the inequities in these courses and the under-representation of women, finding effective ways to measure and improve physics mindset is particularly important in this population if we wish to make physics classrooms more equitable for women and gender minorities. If physics mindset is a useful predictor of learning outcomes, then improving it in physics students may help overall outcomes and equity of outcomes in engineering, physics, and other physical science fields.

6.2 Theoretical background

6.2.1 Intelligence mindset theory

Carol Dweck and her colleagues theorized two types of intelligence mindset—growth and fixed—in the late twentieth century. A growth mindset is one in which intelligence is viewed as something that can be cultivated with effort, like a muscle, whereas a fixed mindset is one in which intelligence is thought to be innate and unchangeable [68]. In the original conception, researchers conceived intelligence mindset as a single continuum in which students varied from having a strong growth mindset at one end of the continuum to having a strong fixed mindset at the other end of the continuum. However, in recent years researchers have used both continuum models [191] and models with separable dimensions in which students can endorse both (or neither) simultaneously [192–194]. The original view holds that as a student ceases to endorse a fixed mindset, they will necessarily endorse a growth mindset [191]. In a two-factor model, it may be possible for a student to endorse neither growth nor fixed beliefs, or they may endorse both types of beliefs. For example, a student might think some basic foundational intelligence or talent is required in addition to seeing value in practice towards further developing intelligence. The mindsets held by a learner are thought to shape how students engage in learning. With a fixed mindset, a student will disengage from or avoid difficult tasks; with a growth mindset, a student will view struggle as an opportunity to learn and gain skills, and therefore will welcome such challenges [69, 70].

The engagement, propensity to attempt challenging problems, and persistence that come with a growth mindset have been linked to positive learning outcomes [72, 73], even after controlling for prior academic achievement [74, 75, 122]. Intelligence mindsets may also play a role in shaping learner self-efficacy [63, 64] and in experiences of anxiety in learning and testing environments [60, 63]. As a result, growth mindsets are not only relevant to improving learning outcomes for all students, but they also may be an important factor in creating equitable classroom environments. For example, having a growth mindset has been linked to greater participation in STEM fields, especially for students from racial and ethnic underrepresented groups [71,195]. Additionally, both students in underrepresented groups and women reported a greater sense of belonging if they endorsed a growth mindset [76].

Growth mindsets can be particularly useful for students as a way to combat stereotype threat. Stereotype threat is "being at risk of confirming, as self-characteristic, a negative stereotype about one's group" [77, p. 797]. For example, a girl or woman taking a math test may feel anxious because of cultural stereotypes that women are not as good at math as men. When such stereotype threats are combined with a fixed mindset, withdrawal from efforts in mathematics can result: the student cannot change their gender, race, or culture, so they may choose to divest from a field that leaves them anxious about representing these identities poorly [77].

Although intelligence mindsets are carried by students into various learning contexts (i.e., have some stability over time and context), they can be malleable through strategic (and relatively brief) interventions with positive results for students' learning outcomes, such as mathematics assessments [196], standardized test outcomes [74], and course grades [75] especially if students are at a high risk of failing a class [197, 198]. However, the effectiveness of both mindset as a predictor of student success as well as the methodology and effectiveness of mindset interventions has been found to vary greatly [199]. For example, only 12% of the interventions included in a recent meta-analysis resulted in significantly greater academic achievement [199], which may make some instructors concerned about their use of class time [200].

The Sisk et al. study [199] explores several potential reasons the effectiveness of these interventions varies, they tend to focus on technical (i.e., if the intervention is online or in-person, the length of the intervention, etc.) differences, which may not be the only aspects of importance. Yeager and Dweck [191] offer more explanations of the varied effectiveness of mindset interventions: first, they show concern about moderation of an intervention's effectiveness at the study level (for example, by length of intervention) rather than the student level (for example, by student gender or socioeconomic status), as it can be difficult to discern the effectiveness of an intervention without simultaneously knowing of methods of the intervention, the students who receive the it, and the larger context the intervention takes place in (e.g., if a growth mindset is supported in the classroom after the intervention). There is also concern about the procedural differences among mindset intervention studies: for example, an intervention that simply explains what a growth mindset is will not be as effective as one that offers students concrete actions to utilize such a mindset [191]. We also hypothesize that some of the varying effectiveness of the interventions may be due to procedural details in the interventions. One possibility is that intervention effectiveness relies on customization to the particular concerns that students have in a particular context. Another possibility is that the focus of the intervention affected its outcome. For example, did the intervention seek only to address the growth mindset but ignore the ability mindset?

Another conceptual divide in mindset research involves beliefs about self versus others. De Castella and Byrne [201] found that Australian high-school students conceptualized intelligence mindsets differently for themselves than for others. They also found that intelligence "self-theory" was a stronger predictor of academic performance and motivation than general intelligence mindsets. Some prior interventions have tried to convince students that people in general can grow their intelligence, leaving relatively untouched the beliefs they have about themselves.

A third issue might also exist in domain-specificity of intelligence mindsets. That is, students might believe that intelligence in general can change through hard work but still have fixed mindsets about particular domains that then more strongly shape how they engage in those particular domains. For example, it was physics-specific mindsets rather than general intelligence mindsets that predicted performance in physics classes [40]. Further, many stereotypes about women and students from underrepresented racial and ethnic groups (for example, Black or Latinx students) are highly domain-specific (e.g., strengths in arts and humanities, weaknesses in math and sciences [56, 202]). Indeed, women in general have higher grades on average in high school and in university [203], so a domain-specific mindset would make more sense as contributing to performance differences in physics courses

6.2.2 Physics intelligence mindsets

There appear to be common views both in society and within the discipline that physics requires a special brilliance. In a study of brilliance beliefs by academic discipline, physics faculty, post-doctoral researchers, and graduate students were more likely to say that physics requires innate talent than those in almost all other fields [59]. Brilliance beliefs are not the same as a fixed mindset, though they work in tandem. If a student thinks raw talent is needed to succeed in a domain (a brilliance belief), and they believe that intelligence is unchangeable (a fixed mindset), then they will see no path to success unless they believe they have innate talent [177]. Indeed, the Leslie et al. study [59] revealed a negative correlation between degree of endorsement in ability beliefs and percentage of PhDs who are women or are from underrepresented racial and ethnic groups, with physics being second highest (after mathematics) among STEM disciplines in field-specific ability beliefs and lowest in percentage of women with doctorates. In a recent study, only half of graduate admissions committees in physics prioritized a growth mindset in their selection process, meaning that they prioritized potential for growth, rather than exclusively seeking out the students with the highest grades and Graduate Record Examinations (GRE) scores [189].

Physics-specific mindset research has just begun in recent years. Interviews show that students [128] and faculty [189] may simultaneously endorse both growth and fixed mindset beliefs, pointing to a need for nuanced measures of mindset. Meanwhile, survey data have provided evidence that students' physics mindsets can be different than their general intelligence mindsets [40]. But an open question regarding physics-specific mindset involves its nature. In particular, does it also separate into independent dimensions of growth and fixed mindsets, with students independently endorsing or denying faxed (fundamental talent) and growth (ability to further improve) aspects? We turn to this issue and potential dimensions of physics mindset in the next section

6.2.3 Dimensions of physics intelligence mindsets

As noted previously, prior research on general intelligence mindsets has gradually transitioned from strict characterization as one continuum (with fixed and growth mindsets on either end [68,75]) to considering a two-factor model measuring endorsement of fxed and growth mindsets separately [192–194], which we denote as growth versus ability because the first label seems to connote the absence of growth rather than the presence of a foundational talent. The primary evidence in favor of treating them separately as two dimensions is psychometric evidence in which a two-factor model produced a better ft to the data. To date, evidence supports a separation of growth and ability dimensions in physics mindsets as well [33], although some researchers have applied a single dimension approach to their data [40, 204].

Another divide which has recently emerged in mindset research is the me versus others distinction. As noted earlier, De Castella and Byrne found that intelligence about the self was the stronger predictor of academic performance [201]. A recent study about physics intelligence mindsets, Kalender et al. found that physics intelligence mindsets divided into four components along the combinations of me versus others and growth versus ability [33]. Although the four components showed some correlations with each other, the best fitting model to the survey data separately measures the four components: My Ability (students' beliefs about their own abilities), My Growth (students' beliefs about their own potential to grow), Others' Ability (students' beliefs about others' abilities), and Others' Growth (students' beliefs about others' potential to grow). Further, the My Ability component was the best predictor of physics course grade, had the largest gender differences, and appeared to largely mediate the effects of gender on grades.

However, the Kalender study uncovered the four physics intelligence mindset components through exploratory quantitative analyses of survey results from a survey that was not designed to separately measure four components of physics intelligence mindsets [33]. There were only one or two survey items for each component's measure, and the items also sometimes differed in other ways across components. In this study, we aim to expand on these results using a larger set of survey items that were specifically designed to measure these four components, allowing for more robust test of the separation into these four components, as well as replicate the three main findings from that study: My Ability was found to be the main predictor of course grades, it was also the component showing the largest gender difference, and it was found to be the only component that mediates the relationships between gender and grades.

6.3 Materials and Methods

6.3.1 Participants

This study takes place at a large, public, urban, predominantly White institution in the northeastern United States. The participants were students enrolled in calculus-based Physics 1 over one semester and across four course sections, each taught by a different instructor. The course covers mechanics and waves, and is taught in a traditional lecture-based format. The N=683 students included in the study were those who completed at least pre- or post-surveys and passed an attention check (a question inserted in the survey that requested students answer "C"). Some (N=39) students were excluded from some portions of this study because they were missing either course grades or prior academic preparation information, though these students were included in the survey validation portion of this study.

Demographic data were acquired from university records. In the student sample, 63% were enrolled in the college of engineering, and virtually all of these engineering students (99%) were in their first semester at the university. The rest of the students were primarily science majors and 59% were from later years at the university. Based upon the data available from the university, women constituted 36% of the student sample. According to university-provided race/ethnicity data, students identified as follows: 73% White, 13% Asian, 7% Hispanic/ Latinx, 4% multiracial, 2% African American/Black, and 1% unspecified.

6.3.2 Measures

6.3.2.1 Demographic information

Students provide demographic information as part of university enrollment. Students were given the binary options “male” and “female” to identify their gender upon entering the university, although this conflates gender and sex [205]. We acknowledge the harm that such data collection practices cause [136], and we are pleased to report that our university has recently switched to collecting gender information using more than binary options. Given the limitations of the data source, the patterns will predominantly reflect patterns of cisgendered women and men. This approach marginalizes non-binary and other gender minority students [136, 158]. However, we use the data collected by the university (i.e., the options provided were female and male while labeled as gender) and refer this variable as “Gender/Sex” in our analysis and results sections [205]. For the quantitative analyses, gender/sex was coded as an indicator variable: women=1, men=0. For race and ethnicity, students were given six options (American Indian or Alaska Native, Asian, Black or African American, Hispanic or Latino, Native Hawaiian/Other Pacific Islander, and White), and students could choose multiple options. Race/ethnicity was only used in description of the sample, and was coded as a series of indicator variables for each major category (i.e., included that racial/ ethnic identity or not).

6.3.2.2 Physics intelligence mindset

We adapted this mindset survey from previously validated surveys [33]. The survey was designed to measure mindsets about self and others, as well as growth- and ability mindsets. Therefore, to be able to separately assess these different aspects of mindset, additional questions were created and some questions were adapted to make the more specific focus more salient. For example, “People can change their intelligence in physics quite a lot by working hard”, becomes “I can change my intelligence in physics quite a lot by working hard.” After the questions were drafted, we used semi-structured cognitive interviews to ensure that students interpreted questions as intended. We conducted 20 one-hour interviews with students who had previously taken physics courses ranging from introductory to graduate-level. Participants were compensated $25. We oversampled on women, given the research focus. A few questions were edited slightly after the interviews, and the questions in the final set were generally interpreted as intended. The final survey had 19 items, each on a four-point Likert scale (Strongly Disagree, Disagree, Agree, Strongly Agree): seven My Ability items, four My Growth items, four Others’ Ability items, and four Others’ Growth items. See Appendix E for the full set of items. For analysis, the four ratings levels were recoded as 1 to 4, with reverse coding for all my ability and others’ ability questions (e.g., questions 5–11 and 16–19). Prior Rasch modeling [108] with this four-point scale for mindset items had found roughly equal psychological distance between levels, justifying use of mean scores [33, 40].

6.3.2.3 Prior academic preparation

Two measures of prior academic preparation were used as control variables in the analyses. High school Grade Point Average (HS GPA) was reported using the weighted 0–5 scale, which is based on the standard 0 (Failing)–4 (A) scale with adjustments for Honors, Advanced Placement and International Baccalaureate courses (all of these programs may offer a "weighted" GPA that adds up to one grade point as a reward to taking advanced courses, which can allow a GPA higher than 4.0). Approximately 1% of students had high school GPAs over 5. They were excluded from the study because their high schools likely used a different grading system. HS GPA is regularly found to predict early undergraduate course performance and is taken to be a measure of general academic skills related to self-regulation, attendance, and putting effort into assignments [138]. Students' Scholastic Achievement Test math (SAT math) scores were used as proxies for mathematical problem-solving skills at the time of university admission. SAT math is one predictor of college performance [138], particularly in quantitative courses like introductory physics [113, 206]. The scores are on a scale of 200–800. We mediated outliers in SAT math by winsorizing [108]. To winsorize the scores, we replaced outliers with values two standard deviations above or below the mean, so that we maintained the direction of the outlier without introducing extreme values. If a student took the American College Testing (ACT) examination, we converted ACT to SAT scores [102]. SAT scores had a negative skew. If a student took a test more than once the school provided the highest section-level score if a student took the SAT and the highest composite score if the student took the ACT. If a student took both tests, we used their SAT score

6.3.2.4 Physics course grade

Physics 1 course grades, the primary course performance measure, were based on the 0–4 scale used at this university, with A=4, B=3, C=2, D=1, F=0 or W (late withdrawal), where the suffixes '+' and '−', respectively, add or subtract 0.25 grade points (e.g., B− =2.75 and B+ =3.25), except for the A+, which is also reported as 4. Each course instructor determined their own grading schemes and there was not a shared departmental exam. However, from examination of syllabi across all sections, course grades were predominantly based upon traditional midterm and final exams, with a smaller portion of the grades based on homework, quizzes, and recitation attendance. Course grades had a negative skew.

6.3.3 Procedures

The surveys were administered to students during recitations associated with the course. The 50 minute recitation sections are mandatory and led by teaching assistants (TAs). The first ("pre") survey was administered on paper during the first or second week of classes, and the final ("post") survey was administered last week of classes. The mindset items were a subset of a larger survey, which took approximately ten minutes to complete. To encourage a high completion rate, students receive either a participation grade or a small amount of extra credit for completing the survey, depending on the instructor's preference. 80% of course enrollees completed the survey at pre and 41% did so at post, reflecting a lower recitation participation at the end of the semester. However, the student sample that completed the survey is very similar to the general population of the course in terms gender/ sex, prior preparation, and course performance, and the students that completed only the pre-survey are similar to those who took both the pre- and post-surveys (see Appendix G). Survey results were collected, de-identified by an honest broker, and then combined with similarly de-identified demographic information and academic history.

6.3.4 Analyses

6.3.4.1 Survey validation

Confirmatory factor analysis (CFA) using the R package "lavaan" was used to both provide quantitative validation of the survey items and to test the proposed conceptual division into four components in terms of growth/ability and myself/others. To evaluate if the model was acceptable, we chose the following standards: standardized factor loadings of each item were all above 0.5 (Kline, 2016, p. 301), a Comparative Fit Index (CFI) and Tucker–Lewis index (TLI) greater than or equal to 0.95 [172], a Root Mean Square Error of Approximation (RMSEA) less than or equal to 0.05 for "good fit" or 0.08 for "fair" fit [107], and a Standardized Root Mean Square Residual (SRMR) less than or equal to 0.06 [172]. The survey was designed to divide items into four categories, but we also explored if a one-factor or two-factor model (i.e., only dividing along one of the aforementioned dimensions, rather than ability/growth and myself/others simultaneously) resulted in a better ft. Poorly fitting items were dropped, and the model was re-evaluated with the remaining items. To create latent variables, we calculated the average score of the questions in each validated category. All the mindset factors are scored from 1 to 4, and are coded such that a high score corresponds with a growth/malleable physics mindset, and a low score corresponds with a fixed mindset. After averaging scores, we winsorized each mindset factor so that outliers were set at a cutoff two standard deviations from the mean of each factor. The resulting variables were used as the mindset measures for the rest of the study.

6.3.4.2 Descriptive statistics

To characterize change in mean attitudes over time, and differences by gender/sex in mean attitudes at pre and post as well as grades, we used Cohen's d to describe the size of the means differences and t-tests to evaluate the statistical robustness of the differences. Cohen's d is considered small if $d \sim 0.2$, medium if d $\sim$ 0.5, and large if d $\sim$ 0.8 [140]. Paired t-tests were used to compare mindset factors between pre and post, while unpaired t-tests were used to compare mindset factors between genders/sexes. Levene's test [108] was implemented to ensure that the homogeneity of variance assumption was met for the unpaired t-tests. We used a significance level of 0.05 in the t-tests and the later regression models as a balance between Type I (falsely rejecting the null hypothesis) and Type II (falsely accepting the null hypothesis) errors [108]. The change-over-time analyses were also done for all instructors separately to check for consistency of the patterns across instructors. Pearson correlations were calculated between the generated latent variables and between the pre- and post-survey scores of the same variable. These correlations can be found in Table 28 and provided information on potential problems of collinearity among predictors in the multiple regressions (e.g., Pearson $r > 0.7$). Further, Pearson correlations allowed us to examine attitude stability over time during this first experience with university-level physics. We also found correlations between mindset factors and course grades as a baseline prediction model.

6.3.4.3 Predicting learning outcomes

Multiple linear regression analysis was used to find partial correlations between mindset factors and grades, controlling for gender/sex and prior preparation. We chose to use regression analysis instead of hierarchical linear modeling because we find the Interclass Correlation Coefficients of motivational factors in these courses are regularly smaller than 0.04 and always smaller than 0.10. Multiple models were tested in order to find which was the best predictor of learning outcomes and show robustness of relationships across model specification. All models used standardized regression coefficients as a measure of effect size. The models were implemented using Stata statistical software [154]. To test the normality of errors, we compared a kernel density estimate of each model's residuals with a normal distribution. Each model had a normal distribution of residuals.

A baseline model predicted grade using only gender/sex, high school GPA, and SAT math scores. Next, we added the mindset variables with the strongest correlation to grade, which can be found in Table 28 one-by-one until all mindset variables were present. All models with significant mindset variables were kept, along with the final model with all variables induced as a robustness test. The regression analyses were repeated with two sets of attitudinal variables: first the scores from the pre-survey, then from the average of pre- and post-survey scores.

The average group included only students who took the survey both times. Average scores were used instead of post-survey scores for two reasons. First, post-survey scores raise the question of causality (did course performance affect mindset or did mindset affect course performance?). Second, the average score is a proxy for students' mindset during the semester, while they were taking the course, rather than after the class. Using average rather than only pre-survey data is particularly important given the sizable changes from pre to post that were observed in several of the attitudinal variables.

6.4 Results

6.4.1 RQ1: What are the components to students' physics intelligence mindsets?

One of the 19 survey items ("I will always be as good at physics as I was in high school.") was removed as a first step because the cognitive interviews show that students did not interpret it as intended. All other survey items appeared to be interpreted as intended. Five additional survey items were removed during the CFA model testing process due to consistently low factor loadings or cross-loading that led to a poor overall model fit. The removed items are indicated in italics on the full survey shown in Appendix E.

Of the four tested models (using all questions in a single factor; splitting by a "growth/ability" dimension; and splitting by a "myself/others" dimension; dividing into four categories in the combination of both dimensions), both two-category models were rejected, as they failed to meet accepted cutoff values for our chosen ft indices.

The third model, which divided survey items into four categories and can be seen in Table 1, had a good model fit. The CFI was 0.95, the TLI was 0.95, the RMSEA was 0.073, and the SRMR was 0.052. Tree of the ft indices—CFI, TLI, and SRMR meet our chosen cutoffs. Our RMSEA meets Browne and Cudeck's [107] $\leq$0.08 guideline for acceptable ft. All standardized factor loadings were above a 0.50 threshold. We named the resulting categories "My Ability" (MA), "My Growth" (MG), "Others' Ability" (OA), and "Others' Growth" (OG). Three categories—MA, MG, and OG—had acceptable values of internal consistency (Cronbach α <0.7), while OA had slightly lower reliability (α =0.68). All four categories had negative skew (e.g., a skew toward a growth mindset). To confirm that these factors held equally well for men and women, we performed measurement invariance testing and found that both weak and strong invariance held for these factors (see Appendix F).

Intercorrelations among the scales are all moderate and positive (after reverse coding of ability), but none are so high as to represent redundant measures. Figure 28 shows that there is also not strong organization of these correlations at the level of the dimensions: while there are some pairwise combinations that are higher, on the whole there are four scales that are all moderately correlated with one another. All four factors show moderate stability over time. Thus, the attitudes that students had at the beginning of the semester could have provided the opportunity to continuously influence student performance and behaviors during the whole semester. However, because there is also significant change, the average attitude held across the semester is likely a better estimate of the relationship of attitudes to performance.

6.4.2 RQ2: Are there gender/sex differences in the different components of students' physics intelligence mindsets?

Table 3 shows descriptive statistics for each measure by gender/sex at pre and post. On the pre-surveys, men and women have nearly identical and high scores in the My Growth, Others' Ability, and Others' Growth categories. That is, in general most students have growth rather than fixed mindsets, particularly when considering others.

The only pre-survey category with a significant gender/sex difference is My Ability. In this category the gender/sex difference has a medium effect size with men having higher scores than women (i.e., women were more likely than men to believe that natural ability is important for themselves to succeed in physics). There were also gender/sex differences in prior academic performance. As seen in Table 3, women tend to have higher high school GPAs than men in our sample, but lower SAT math scores. Both of these differences had relatively small effect sizes and both populations generally had high scores (i.e., were generally well prepared for challenging academic work). Thus, the lower average course grades for women (see Table 3) are somewhat surprising from an academic preparation perspective. Note, we cannot assume men's higher SAT math scores directly translate into higher grades in math-intensive courses. There is no similar gendered grade difference in Calculus 1, which this population often takes in tandem with Physics 1 [14, 39]. Instead, factors other than academic preparation are likely at play.

6.4.2.1 RQ3: If there are differences in the components of students' physics intelligence mindsets, do the differences grow or decline from the beginning to the end of their first university-level physics course?

By the end of the semester, there were moderate-to-large gender/sex differences in all four mindset constructs, and all gender/sex contrasts became statistically significant. Thus, following their first experience in university-level physics, women were more likely than men to believe that natural ability is important to succeed in physics for both themselves and others. This change in gender/sex differences reflects moderate-to-large declines in attitudes in women but only small declines in men, on average (see Figure 1). This suggests that classroom experiences that influenced student mindsets affected men and women differently. Trends were similar across instructors, though some results were non-significant when calculated for individual instructors' classes, due to low sample size.

6.4.2.2 RQ4: Do any of the mindset factors from RQ1 predict course grade?

We conducted multiple regression analysis to find which of the four mindset factors best predicted physics course grade (see Table 4). Models 1–3 used only pre-survey results, while Models 4–6 used the mean of pre- and post-survey mindset scores (because of the large changes in mindset across the semester). In Model 1, only gender, SAT math scores, and HS GPA are included as predictors and all three were statistically significant. This model shows that women have lower Physics 1 grades than men when controlling for prior academic preparation, formally establishing that other factors are needed to account for gender/sex differences in course performance.

Model 2 includes My Ability (MA) as a fourth predictor, the single strongest correlate of grades. Here pre-survey MA is a significant predictor beyond academic preparation. Its addition weakens the relationship between gender/sex and Physics 1 grade, though gender/sex remains significant. Additionally, Model 2 has a small increase in adjusted R-squared compared to Model 1. This means that Model 2 explains more of the variance in course grades than Model 1, while penalizing for non-signifcant predictors [108].

Model 3 adds the remaining pre-survey mindset factors: MG, OA, and OG. None of the newly added factors are statistically signifcant, and their addition leaves fully intact or slightly strengthens the predictive power of the other predictors, suggesting robust relationship estimates. The predictive power of gender/sex decreases slightly. Variance Inflation Factors (VIFs) for every variable in Models 1–3 were below our cutoff of 2.0, which indicates that our models are not skewed by multicollinearity, even in the case of the different mindset factors that were also moderately correlated with one another.

Models 4–6 are focused on the sample that completed both pre and post to unpack the predictive role of average attitudes across pre and post. Model 4 is identical to Model 1, but now providing the baseline model for the reduced sample set. The parameter values are similar in approximate magnitude as those of Model 1, although the SAT estimate is smaller and the gender/sex estimate is larger.

Model 5 adds average MA as a predictor. Average MA has more than twice the predictive power of pre-MA, and the gender/sex estimate decreases in size by 40%. Model 6 introduces the remaining average mindset factors, none of which are statistically significant predictors, similar to the findings of Model 3. There are no major changes in the predictive power of MA, HS GPA, or SAT math from Model 5 to Model 6, again suggesting robust relationship estimates and that MA in particular was the most likely mediator of gender/sex differences in grades among the mindset factors.

In Models 4–6, VIFs are mostly below the cutof of 2.0, except for MG (VIF=3.08) and MA (VIF=3.19) in Model 6. MA and MG are often conceptualized as a single factor [207] because they have substantial intercorrelations [194] as in our analysis (see Table 3). However, the robustness of the pattern of regression estimates and much lower predictive power of MG across models supports the focus on MA as the key predictor of student performance. Although there was analytic support for treating the Likert ratings as continuous predictors, some skew in the distributions did occur. However, regression results were very similar when binary mindset variables (i.e., 1 for strong endorsement of growth mindset/strong rejection of fxed mindset; 0 otherwise) were used instead of means based upon 1–4 codings. Most importantly, MA was the strongest predictor of grade among the mindset factors.

6.5 Discussion

6.5.1 RQ1: What are the components to students' physics intelligence mindsets?

The current study strongly replicated the exploratory findings of Kalender et al. [33] using a survey instrument designed to specifically test for the four components: My Ability (MA), My Growth (MG), Others' Ability (OA), and Others' Growth (OG). It also builds upon the work of De Castella and Byrne [201], who found an empirical separation of my versus others' mindset factors, along with a number of other studies that found support for a divide along the ability/effort dimensions [192–194]. The four components were only moderately correlated with one another (∼25% shared variance at pre) and were separable in CFA models. Further, the My Ability factor showed different patterns related to RQ2 and 3. In sum, there was support for separating our four components both from psychometric analyses and empirical phenomena.

6.5.2 RQ2: Are there gender/sex differences in the different components of students' physics intelligence mindsets?

At the start of the semester, there were no gender/sex differences in My Growth, Others' Growth, or Others' Ability. However, there was an initial (moderately sized) gender/sex difference in My Ability even among this relatively selective set of students who have opted into engineering and physical science pathways. That is, women in this context were more likely than men to believe that physics requires innate ability and that they, in particular, did not possess that ability. However, by the end of the semester, all four mindset categories showed significant gender/sex differences, and sometimes large differences.

6.5.3 RQ3: If there are differences in the components of students' physics intelligence mindsets, do the differences grow or decline from the beginning to the end of their first university-level physics course?

Both self-theory mindset factors (My Ability and My Growth) significantly decreased (i.e., mindsets became less growth-oriented and more fixed) for men from the start to the end of the semester, while all intelligence mindset factors significantly decreased for women. In addition to decreasing all mindset factors for students regardless of gender, the courses also created or contributed to a gender-based inequity in physics intelligence mindsets. These results add to research showing that women in physics courses also have other forms of lower average motivational characteristics, such as self-efficacy and sense of belonging, than do men, even in highly self-selected pathways [11, 12, 40]. Such differences may come from general messages about the discipline. In physics, and a few other fields, success is often viewed as a result of brilliance [59] and women may receive fewer messages that they are brilliant and can thus succeed in physics. Such differences may also come from differential experience. In the US, women make up less than a third of students who take advanced (Physics 2 or AP Physics C) high school physics [66].

6.5.4 RQ4: Do any of the mindset factors from RQ1 predict course grade?

Despite having only small differences in SAT Math scores and compensatory strengths in HS GPA, women had lower grades in this physics course. Mindset differences, especially related to My Ability, offer a partial explanation for this phenomenon. Based on our regression models, My Growth, Others' Ability and Others' Growth did not predict course grade, while both pre- and average-My Ability did. Note, however, that less than half the grade gender/sex difference was explained by the My Ability component. It may be that other motivational factors, such as self-efficacy [11, 12, 40, 41, 65, 188], were also important contributors to students' final grades. Alternatively, differences in the learning environment, such as microaggressions by peers, TAs, and instructors, or differential levels of support, may also have played an important role in the differential learning outcomes [57, 208].

Because physics self-mindset is a predictor of Physics 1 grade, finding a way to increase My Ability beliefs may mitigate gendered grade differences. In this population (primarily engineering students) women are more likely to leave the major due to concerns about low grades than men are, even when they have an A or B average [115], so enhancing women's My Ability beliefs may increase retention. Importantly, average My Ability is a stronger predictor of course grade than pre-My Ability.

Thus, educators have an opportunity to intervene and potentially improve grades and cultivate growth mindsets, especially since (from RQ1) mindset self-theory appears to be malleable during this time period. If self-mindset is simultaneously more malleable and has a stronger correlation to learning outcomes, than mindset interventions in this context should focus on students' individual experiences or the experiences of people they can relate to (for example: [132, 209, 210]), rather than activities that focus on teaching students about the brain's general ability to change and grow (for example [74, 75]). The latter approach appears to be well-suited to students who hold a general fixed mindset. However, it may not be useful to students who endorse a general growth mindset but a fixed self-theory. In addition to showing students that changing one's intelligence is possible, we must show them that they can change their own intelligence.

6.5.5 Teaching implications

For instructors who want to help students abandon fixed mindsets, student-level interventions can be valuable. It is especially important in disciplines like physics, where endorsing a fixed mindset is common [59], that instructors clearly state that hard work and effort are necessary for success, not innate ability. Providing opportunities for self-refection about times that students improved their abilities, or sharing stories of a diverse (so that all students in the class will be able to relate to some examples) range of people that overcame academic challenges may also help students develop a growth mindset and improve academic outcomes [64, 132].

Instructor-facing interventions can be useful, too. Discipline-wide mindset beliefs can predict the diversity of graduate programs [59], but do not predict student course achievement as well as the mindset of instructors do [211]. Instructors with fixed mindsets tend to have low expectations of students they believe lack natural talent, which can lead instructors to give easier assignments or encourage students to drop difficult classes because of presumed low ability [212]. Instructors with growth mindsets encourage students to accept mistakes and failures as a part of a normal learning process, congratulate persistence, and praise effort rather than intelligence when students succeed [209, 212]. Instructors with growth mindsets are also more likely to implement active learning in their courses [213]. Students report decreased interest in courses, as well as more concerns over fair treatment and low grades if their instructor had a fixed mindset as opposed to a growth mindset, and this effect was larger for women than men in the study [214].

6.6 Limitations and future directions

The primary goal of this research was to identify which physics intelligence mindsets participate in important empirical phenomena: changing after instruction, predictive of course grades, and potentially explaining gender/sex differences in course grades. However, it is important to acknowledge that the analyses were fundamentally correlational in nature. The causal relationship of physics intelligence mindsets would need to be further supported through intervention studies. The established benefits of other mindset interventions (e.g., Felder et al. 1995 [215]) suggest such a causal link is plausible. Further, the more specific physics intelligence mindset factor most directly associated with course grades (My Ability) suggests a new focus for mindset interventions that could have even larger effects.

A second set of concerns relate to generalizability of the findings. Because the studied institution is predominantly white, we were unable to study if mindset beliefs differ or predict grades differently for students of different racial/ethnic backgrounds due to low sample size. Although the findings were stable across the instructors in the study, a broader set of instructional contexts should also be examined. It may be that other instructional formats (e.g., with well-supported group-work) or more gender-balanced courses would produce smaller declines in physics intelligence mindsets [120]. However, due to regular replications of related research [11, 41, 65], we believe our results are likely to translate directly to other large state universities. Results from different contexts, like such as liberal arts and community colleges, as well as schools that are much more or less selective than our institution, should be examined.

Due to the focus on gender in this study, future research should also explicitly include students who fall outside of the binary gender/sex categories included here, as well as transgender students who may not have their gender accurately recorded by the university. Though this university recently began to include more sex/gender options for students, qualitative studies may be more appropriate to understand mindset in these marginalized populations until student samples are large enough to be meaningful in quantitative analysis.

Another dimension of generalization relates to other disciplines. This study focused on gender/sex and physics mindsets because women are an underrepresented group in physics. Because the intelligence mindsets are likely important in other STEM disciplines, generalizability should be tested in other fields, especially where women are more equitably represented (e.g., biology and chemistry). The patterns across disciplines will provide important clues into the mechanisms that produce these effects.

6.7 Conclusions

Mindset research has recently garnered attention in the physics context. This study shows that intelligence mindset can be divided into four factors: My Ability, My Growth, Others' Ability, and Others' Growth. Previous work studying mindset has divided along either by growth/ability or me/others categories, but rarely simultaneously. However, qualitative studies in physics have called for a more nuanced measurement of mindset than most surveys allow; these four categories are a step in that direction. Next, this work reveals that gender/sex differences are more pronounced in the "My" categories than the "Others" categories, and these differences are developed or exacerbated from the start to the end of an introductory physics course. These results show that women's and men's intelligence mindsets are affected differently by the classroom environment, and future studies may find this useful when developing new interventions or teaching methods aimed at helping students develop growth mindsets. Finally, we find that My Ability is the only mindset factor that predicts course grade. This information may be useful to target mindset interventions to student beliefs. A student who believes nobody can become more intelligent through hard work has very different needs than one who believes that most people can become more intelligent but that they personally lack the ability to do so.

Table 24: Survey items included in the study and standardized factor loadings for pre and post surveys. N=781

Construct name or Item test	λ
My Growth ($\alpha = 0.84$)	
I can become even better at solving physics problems through hard work	0.76
I am capable of really understanding physics if I work hard	0.83
I can change my intelligence in physics quite a lot by working hard	0.82
My Ability ($\alpha = 0.84$)	
Even if I were to pend a lot of time working on difficult physics problems, I cannot develop my intelligence in physics further	0.64
I won't get better at physics if I try harder	0.64
I could never excel in physics because I do not have what it takes to be a physics person	0.87
I could never become really good at physics even if I were to work hard because I don't have natural ability	0.87
Others' Growth ($\alpha = 0.84$)	
People can change their intelligence in physics quite a lot by working hard	0.82
If people were to spend a lot of time working on difficult physics problems, they could develop their intelligence in physics quite a bit	0.82
People can become good at solving physics problems through hard work	0.77
Others' Ability ($\alpha = 0.68$)	
Only a few specially qualified people are capable of really understanding physics	0.67
To really excel in physics, people need to have a natural ability in physics	0.73
If a student were to often make mistakes on physics assignments and exams, I would think that maybe they are just not smart enough to excel in physics	0.55

Table 25: Pearson correlations between each mindset construct as well as physics 1 course grade. The following abbreviations are used: My Ability (MA), My Growth (MG), Others' Ability (OA), and Others' Growth (OG). $p < 0.001$ unless otherwise noted by: $^{*} = p < 0.05$, $^{**} = p < 0.01$, and ^{ns} = not statistically significant.

	Pre				Post			
	MG	MA	OG	OA	MG	MA	OG	OA
MG Pre					0.34			
MA Pre	0.52					0.44		
OG Pre	0.51	0.43					0.38	
OA Pre	0.28	0.45	0.33					0.33
MG Post								
MA Post					0.67			
OG Post					0.69	0.57		
OA Post					0.43	0.65	0.44	

7.0 Bioscience student' internalized mindsets predict grades and reveal gender inequities in physics courses

7.1 Introduction and theoretical framework

For decades, physics departments have struggled to recruit and retain women [1–3] and generally many women in the broader STEM workforce have a negative view of physics [216]. In response, researchers have dedicated effort into improving gender equity and diversity (for example, see [34, 132–135, 217, 218]) of physics departments and classrooms. Some of their research has focused on gender differences in motivational beliefs that arise from negative messages in prior and current classrooms as well as broader society. For example, researchers have found that gender differences in physics-specific motivational beliefs (such as physics self-efficacy, perceived recognition from instructors, and intelligence mindset) may account for some of the differences in physics performance and persistence between women and men [10,16,23,33,128–131]. Other studies also posit that societal stereotypes and biases about who belongs in and can excel in physics also may explain some of these gender differences [55–58], either via messages from media, family, and friends or as the cause of negative messages voiced by instructors, TAs, and classmates [10, 130, 131].

Much of the research about motivational beliefs, performance, and equity in introductory physics courses has focused on courses for students pursuing engineering and physical science majors, rather than for bioscience and health-related majors. These courses in the US often differ in gender/sex makeup: most students in courses for engineering and physical science students are men, but most students in courses for bioscience students are women, similar to the higher participation rate of male students in calculus-based vs. algebra-based AP physics courses [219]. Some research suggests being a numerical minority in a classroom has negative motivational consequences [220]. On the other hand, negative prior experiences with physics may continue to produce negative attitudes towards physics even when one is not a numerical minority. Other prior research has found that even in physics courses in which women are not underrepresented, men tend to have higher grades and physics-specific motivational beliefs in physics courses than women [10, 16, 23, 33, 34, 128–134, 158]. For example, women tend to have lower physics self-efficacy (one's belief in their capability to succeed at an activity or subject [127]) than men with the same grades in courses for engineering and physical science students as well as courses for bioscience students [23, 41]. However, one motivational belief that has not been studied in the context of bioscience majors taking introductory physics is physics intelligence mindset, which may be particularly skewed towards fixed mindsets among students choosing majors and career paths that involve relatively little physics.

More broadly, intelligence mindset describes a person's views about the nature of intelligence, and was originally conceptualized on a spectrum [68]. On one end of this spectrum is a growth mindset, in which intelligence is believed to be cultivated with effort and can be developed over time [68]. On the other end is a fixed mindset, in which intelligence is believed to be innate and unchangeable [68]. The study of domain-specific intelligence mindset has gained popularity in recent years [16,33,128]. This is because the mindset for a discipline can be different from a general intelligence mindset and because domain-specific mindsets tend to be more predictive of student performance in that discipline [16, 33, 41].

In prior work, we developed a new, physics-specific tool to measure intelligence mindsets [16, 33], which has been previously used to investigate mindset beliefs of students in physics courses aimed at engineering and physical science majors, but has not yet been used for students in physics courses for bioscience and health-related majors. In this study, we aim to investigate the nature of physics-specific mindsets for this latter group, as well as whether physics intelligence mindsets change from the beginning to the end of the course, differ by gender, or can predict learning outcomes.

7.1.1 Intelligence Mindset Theory

Intelligence mindset theory posits that there are two broad beliefs about intelligence and how it is formed: growth mindsets and fixed mindsets. A growth mindset is one in which intelligence is viewed as something that can be cultivated with effort, like a muscle, whereas a fixed mindset is one in which intelligence is thought to be innate and unchangeable [68]. Mindset beliefs have implications for how people engage with challenges faced while learning. Students with fixed mindsets tend to disengage from or avoid difficult tasks, and tend to view struggle as a sign that they are not smart enough to succeed, rather than a normal part of learning [68, 70, 71]. On the other hand, students with growth mindsets tend to welcome challenges and view them as an opportunity to learn and improve their abilities [69, 70].

Intelligence mindsets are a useful focus for educational research because of their important role in influencing student learning behaviors but also because relatively brief interventions have been found to successfully change student mindsets for months and even years later. Focusing on their role in student learning behaviors, growth mindsets have been linked to positive learning outcomes even after controlling for prior academic achievement because they can increase students' engagement, propensity to attempt challenging problems, and persistence [68, 72–75]. Further, intelligence mindsets often vary by gender and race/ethnicity, and these relationships have been argued to be an important pathway by which inequity of learning outcomes and participation in STEM occur [71, 195]. Strong growth mindset beliefs can lead to a greater sense of belonging for both women and students from other underrepresented groups [76].

Turning to interventions focused on student intelligence mindsets, a number of brief interventions have been tested in middle school, high school, and university contexts. Several of these interventions have successfully changed students' reported mindsets [75, 132, 198, 221] and improved students' learning outcomes [74, 75, 196]. These interventions have tended to be especially effective for students at high risk of failing a class [197, 198].

Despite some well publicized successes with some interventions, a recent meta-analysis by Sisk et al. [199] revealed that the effectiveness of mindset interventions varies significantly, with only 12% of included interventions significantly improving academic achievement. One possible factor that could determine the effectiveness of a mindset intervention is the demographic groups a student belongs to. An intervention may be more effective for women or low-income students than for men or high-income students [191]. Indeed it is important to examine which groups experience low growth mindsets or high ability mindsets to understand which groups are likely to be helped by a mindset intervention. However, Sisk et al. also raised concerns about mindset's ability to predict learning outcomes in particular contexts. We argue (see below) that general intelligence mindsets may not be as important as discipline specific mindsets for participation and learning outcomes within disciplines, especially in disciplines like physics for which there are especially strong stereotypes about brilliance [59].

7.1.2 Dimensions of Intelligence Mindset

Researchers initially viewed intelligence mindset as a single continuum in which a growth and fixed mindset sit on either end [68]. However, interviews show that students may simultaneously have some growth mindset beliefs and and fixed mindset beliefs, pointing to a need for more nuanced dimensional measures of mindset [128, 189]. Technically speaking, the former approach is a "one-factor" model, while the latter is a "multi-factor" model. Though the one-factor approach is still popular [42, 191, 204], there is a growing body of work that uses separable growth and fixed mindset dimensions [16, 33, 192–194]. For example, in a two-factor model, a student might report both come growth mindset and some fixed mindset beliefs. Such a student may understand that practice and hard work are necessary to excel in physics. However, that student may also believe a base level of ability is also needed and feel disempowered if they think that they do not personally posses that "necessary" talent or ability to excel.

Another conceptual divide in mindset research involves beliefs about self versus others. One study [201] found that high-school students conceptualized intelligence mindsets differently for themselves than for others. They also found that intelligence "self-theory" was a stronger predictor of academic performance than general intelligence mindsets. As noted in the next section, similar patterns were recently found with self vs. other physics mindsets.

We aim to investigate if students had separable beliefs about growth and fixed mindsets, as well as if they held different mindset beliefs about themselves versus others. If student mindsets are separable along these divides, then there is an opportunity to learn which more specific mindsets are particularly important for learning outcomes or especially associated with gender differences. Those findings in turn would better enable targeted interventions.

7.1.3 Physics Intelligence Mindsets

Students may have different mindset beliefs in different domains and contexts. For example, they may believe that intelligence in general can change through hard work or that they in general have enough intelligence for most situations, but still have fixed mindsets about particular domains with especially strong stereotypes of innate brilliance such as physics. Physics-specific mindset research is relatively new [16, 33, 42, 128]. One of these first studies found that physics-specific mindsets are both different from (via a factor analysis) and a better predictor of physics learning outcomes than general intelligence mindsets [42].

Further, many stereotypes about women and intelligence are domain-specific. For example, women are perceived to have strengths in the arts and humanities and weaknesses in math and the sciences [56, 202]. Physics in particular is a field with particularly strong stereotypes and biases about who belongs in and who can excel in the domain [55, 56, 177]. Both the general public [55] and working physicists [59] believe that success in physics requires innate talent or brilliance and societal narratives about talent and brilliance tend to ascribe these traits to boys and men [57, 177, 222]. Parents of girls are less likely to believe their child could succeed in a career that requires mathematical ability [67, 97]. Boys are more likely than girls to receive positive recognition from their science instructors, including in physics courses [9, 28, 131]. Finally, there is evidence that physics intelligence mindsets become more fixed after taking a physics course, especially for women [42].

Recent research supports a four-way division of physics intelligence mindsets, and finds that one of the four physics-specific mindsets was especially predictive of introductory physics course grades in the male-dominated courses for physical science and engineering majors [16, 33]. In particular, Kalander et al. were the first to find that physics intelligence mindsets can be divided into four dimensions along the combinations of me versus others and growth versus ability and the best fitting model to the survey data separately measures the four factors: My Ability (students' beliefs about their own abilities), My Growth (students' beliefs about their own potential to grow), Others' Ability (students' beliefs about others' abilities), and Others' Growth (students' beliefs about others' potential to grow) [33].

However, the Kalender et al. study uncovered these four mindset factors using a survey that was not specifically designed to measure four dimensions of physics intelligence mindset (i.e., had too few items per dimension) because this was not the original conception that drove the design of that survey instrument [33]. Malespina et al. then built upon this work in the same context by expanding the number of survey items and designing their structure to directly map onto the four hypotheses components and was able to replicate the original findings [16]. Further, both studies (each conducted in the calculus-based context for engineering and physical science majors) found that My Ability was the best predictor of physics course grade, had the largest gender differences, and appeared to largely mediate the effects of gender on grades.

7.1.4 Research Questions

Here the same survey items from the Malespina et al. study are used in a new context: introductory physics for bioscience majors [16]. The survey aims to pinpoint specific mindset beliefs (such as if a student holds different mindset beliefs for themselves versus their peers). Additionally, the measure is context-specific to physics. We will also examine whether mindset beliefs predict learning outcomes differently for men and women, as suggested by Yeager and Dweck's [191]. In addition, this research will investigate whether student grades are predicted by mindset across the full range of possible mindset levels or whether there are threshold effects such that mindset differences only matter at the high or low end. Here, we use "low", "medium", and "high" threshold values to measure student mindset. Though these thresholds are specific to the instrument used in this study, such threshold effects could help investigate which courses and students are most in need of intervention. For example, if outcomes for low and medium mindset values are similar, then it would be important to prioritize high scores for students through interventions and other means. We aim to answer the following research questions for students in introductory physics courses for bioscience majors at a large research university:

RQ1. Do physics intelligence mindsets organize into four factors (My Ability, My Growth, Others' Ability, and Others' Growth) as they did for students enrolled in physics for engineering and physical science majors?

RQ2. a. Are there overall gender/sex differences in the means or distributions (in low, medium, and high categories) of students' physics intelligence mindset beliefs?

b. Are gender/sex differences in the means or distributions of students' physics intelligence mindset beliefs especially localized to particular dimensions?

c. Do gender/sex differences grow or decline during students' first university-level physics course?

RQ3. Do any of the mindset dimensions predict course grade and is the predictive relationship linear?

If the findings replicate what was found in the male-dominated introductory physics courses for physical science and engineering majors, then we expect: 1) four dimensions (My Growth, My Ability, Others' Growth, Others' Ability); 2) men will have higher mindset scores than women, especially for My Ability beliefs, and gender differences in all of the mindset factors will grow over time; and 3) My Ability is the best predictor of grade.

7.2 Methodology

7.2.1 Participants and Procedures

We collected survey data at the beginning and end of the semester. Participants were students enrolled in a physics 1 course for bioscience and health-related majors. At this institution, introductory physics courses for bioscience majors are algebra-based, while courses for physical science and engineering majors are calculus-based. The physics 1 course primarily covered mechanics, though both thermodynamics and and waves were also included. Faculty taught the course in a traditional lecture-based format alongside smaller-sized recitations taught by teaching assistants in which students work collaboratively on physics problems. The student sample involved three different sections taught by three different instructors in one semester.

Surveys were handed out and collected by the teaching assistants in the first and last recitation class of a semester. Students were given course credit or extra credit for completing the survey, depending on the instructor's preference. The completion rate was 83% ($N = 547$) for the pre test and 78% ($N = 500$) for the post test. We focused upon the 428 students who took both surveys so we could observe students' change in motivational beliefs over time; however, similar findings were obtained when using the full set of respondents. An additional 9 students were excluded from the study due to missing demographic information or receiving an “incomplete” grade in the course. The final number of students in the presented analyses was 419.

Based upon institutional data, the longitudinally matched sample was 66% women (compared to 62% for all enrolled students), which is typical for introductory physics for bioscience majors courses at this institution. We note that response rates were slightly different by gender (Pre/Post response rates were for 85%/79% for women and 79%/70% for men). Students at this predominantly white institution (PWI) identified with the following races/ethnicities: 68% White, 19% Asian, 3% Hispanic/Latinx, 5% multiracial, and 5% African American/Black. This course is taken almost exclusively by students intending to pursue postgraduate work in the health fields (especially medicine). Most students were in their second (13%) or third (65%) year of university.

This research was carried out in accordance with the principles outlined in this institution’s Institutional Review Board ethical policy, and de-identified demographic data were provided through university records. For some variables, such as high school GPA, this approach allows us to rely on records that may be more accurate than students’ memories. However, it limits other measures such as student sex/gender, which students could only report as “male” or “female”. We acknowledge the harm that collecting data this way can cause [136, 158]. This institution recently began to implement more inclusive sex and gender reporting methods for students, which we plan to use once student samples are large enough to be meaningful in quantitative analysis.

7.2.2 Measures

7.2.2.1 Physics intelligence mindset

We adapted this mindset survey from previously validated surveys [16, 33, 42]. The survey was designed to measure mindsets across a self vs. others dimension, as across a growth vs. ability dimension. Initially, there were 19 items in the mindset survey, which can be found in Appendix H.

After the questions were drafted, we conducted 20 one-hour semi-structured cognitive interviews to ensure that students interpreted questions as intended. Participants were students who had previously taken physics courses ranging from introductory to graduate-level. One of the 19 survey items ("I will always be as good at physics as I was in high school.") was removed because the cognitive interviews indicated that students did not interpret it as intended [16].

This survey was designed to have four separable mindset beliefs based upon the combinations of those two dimensions [16,33]: My Ability, My Growth, Others' Ability, and Others' Growth. The items were distributed across mindsets as follows: six My Ability items, and four items each for the three other constructs. See Appendix H for the full set of items. Each item used a common set of four response options (Strongly Disagree, Disagree, Agree, Strongly Agree). Responses were correspondingly coded as 1 to 4, with reverse coding for all My Ability and Others' Ability questions such that higher values corresponded to mindsets hypothesized to support learning.

7.2.2.2 Prior academic preparation

High school Grade Point Average (HS GPA) was reported using the weighted 0–5 scale, which is based on the standard 0 (Failing)–4 (A) scale with adjustments for Honors, Advanced Placement and International Baccalaureate courses (all of these programs may offer a “weighted” GPA that adds up to one or two grade points as a reward to taking advanced courses, which can allow a GPA higher than 4.0). High School GPA is taken as a measure of general academic skills and generally is a strong predictor of early undergraduate course performance [138]

Students’ Scholastic Achievement Test math (SAT math) scores are on a scale of 200–800 and were used as a predictor of performance on high-stakes assessments involving mathematical problem-solving (e.g., physics exams) [113, 138, 139]. If a student took the American College Testing (ACT) examination, we converted ACT to SAT scores [102]. If a student took a test more than once the school provided the highest section-level score for the SAT and the highest composite score for the ACT. If a student took both ACT and SAT tests, we used their SAT score.

7.2.2.3 Course Grade

Course grades were based on the 0-4 scale used at our university, with A = 4, B = 3, C = 2, D = 1, F = 0 or W (late withdrawal), where the suffixes ‘+’ and ‘-’, respectively, add or subtract 0.25 grade points (e.g., B- = 2.75 and B+ = 3.25), except for the A+, which is reported as 4.

7.2.3 Analysis

7.2.3.1 Survey Validation

Confirmatory factor analysis (CFA) using the R package “lavaan” was used to provide quantitative validation for whether the survey items fit the proposed four mindset constructs. To evaluate whether the model was acceptable, we chose the following standards: standardized factor loadings of each item were greater than 0.5 [173, p.301], a Comparative Fit Index (CFI) and Tucker–Lewis index (TLI) greater than 0.90, a Root Mean Square Error of Approximation (RMSEA) less than or equal to 0.08, and a Standardized Root Mean Square Residual (SRMR) less than or equal to 0.08 [223].

We first investigated whether the conceptual division into four components in terms of growth/ability and myself/others was replicated in this course context. In particular, in addition to testing the fit of the model based upon the four categories, other models were also evaluated based upon other approaches to intelligence mindset. A one-factor model in which all items were included in a single construct was tested and rejected due to poor model fit. Two-factor models were also evaluated: one model divided items that asked about the self and others, and the other divided questions that asked about growth and ability. Both models were rejected due to poor model fits. The four-factor model resulted in the best overall model fits.

After deciding on a four-factor model, the item with the lowest factor loading was dropped, and the model was iteratively re-evaluated with the remaining items. Items were dropped as long as fit indices improved or remained consistent and each factor had at least three items. This process produced a robust model while eliminating excess variables. After determining the items to include, we calculated Cronbach’s

α, a measure of internal consistency between items within a construct. A generally accepted value for Cronbach's α is between between 0.70 and 0.90 [108].

To create latent variables, we calculated the mean score of the questions in each validated category using the reduced set of twelve survey items. As a reminder, all the mindset dimensions are scored from 1 to 4, and are coded such that a high score corresponds to agreeing strongly with growth/malleable physics mindset beliefs or disagreeing with fixed/ability mindset beliefs. We used mean scores for constructs because prior Rasch modeling [108] with this four-point scale for mindset items had found roughly equal psychological distance between levels [33] and because the correlation between simple mean scores and Rasch-adjusted person estimates are very high (e.g., usually above .99).

We also tested different levels of measurement invariance in the final CFA model to make sure the survey items functioned equally across gender groups given the focus of the current study. In each step, we fixed different elements of the model to equality across gender and compared the results to the previous step using the likelihood ratio test [173]. We did not find any statistically significant moderation by gender, supporting the use of mean scales scores in analyses of gender differences.

After completing the CFAs to determine the mindset scales, we addressed outliers in all mindset scale values (as well as in SAT/ACT math, course grade, and high school GPA) by winsorizing [108]. To winsorize the scores, we replaced outliers with values two standard deviations above or below the mean, so that we maintained the direction of the outlier without introducing extreme values that produce poor performance in the regression models.

Pearson correlations were calculated between the generated latent variables within a time point to provide information on potential problems of collinearity among predictors in the multiple regressions (e.g., Pearson $r > 0.70$). Further, pre-post Pearson correlations for each attitude were used to examine attitude stability over time: pre-post correlations below 0.3 would indicate low stability, correlations above 0.8 would indicated high stability, and intermediate values would indicated moderate stability. We also calculated Pearson correlations between each mindset dimension and course grades as a baseline prediction model.

7.2.3.2 Descriptive Statistics

To analyze gender differences on all measures, we calculated means and standard deviations by gender and then we compared men and women's scores using unpaired t-tests to measure statistical significance of the differences [108] and Cohen's d to measure the size of the difference [140]. Cohen's d is calculated using:

$$d = \frac{\mu_1 - \mu_2}{\sqrt{(\sigma_1^2 + \sigma_2^2)/2}}, \tag{2}$$

where μ_1 and μ_2 are the mean values of each group and σ_1 and σ_2 are the standard deviations of each group [140]. Group one was women and group two was men. Cohen's d is considered small if $d \sim 0.2$, medium if $d \sim 0.5$, and large if $d \sim 0.8$ [140]. Levene's test was implemented to ensure that the homogeneity of variance assumption was met for the unpaired t-tests [108].

Similarly, to compare students' mindset scores from pre to post, paired t-tests [108] and Cohen's d effect size measures were also used. The change-over-time analyses were also conducted for all instructors separately to check for consistency of the patterns across instructors. Trends were generally similar between instructors. One instructor's class did not have statistically significant mindset decreases in all constructs for men, though the decrease was similar in magnitude to other instructors. This may be due to small class size, as there were only 22 men in that group of participants.

Finally, we divided students into groups that reported "low" (< 2.5), "medium" ($2.5 - 3.5$), and "high" (≥ 3.5) on the 1-to-4 scales (after recoding) for each mindset dimension. The specific thresholds were selected given the distribution of the data, as it was rare for students to select the lowest values for each survey ite. We divided students into categories for two reasons. First, for instructors with large class sizes, strategically dividing students into groups with low, medium, or high mindset scores may be easier to manage than placing students into groups based upon a continuum. Second, analyzing the data this way provides a test of the linearity assumption in the regression analyses. Third, if effects were non linear, this could shape the scale of interventions that would be needed (e.g., for moving students from low all the way to high).

7.2.3.3 Predicting Learning Outcomes

First, multiple linear regression analysis was used to find partial correlations between mindset components and grades, controlling for gender/sex and prior academic preparation. For the quantitative analyses, gender/sex was coded as an indicator variable: women=1, men=0.

Regression analysis was chosen over hierarchical linear modeling because the Interclass Correlation Coefficients of the motivational measure data in these larger lecture courses are adequately small (always < 0.10 and regularly < 0.04). Multiple models were evaluated in order to find which was the best predictor of learning outcomes and show robustness of relationships across model specification. All models used standardized regression coefficients as a measure of effect size. The models were implemented using the regress command in Stata [154]. To test the normality of errors, we compared a kernel density estimate of each model's residuals with a normal distribution. Each model had a normal distribution of residuals. We also ensured that predictor effects were not miss-estimated due to multicollinearity by implementing a Variance Inflation Factor cutoff of 2.0 for each model.

The regression models were built incrementally to assess the robustness of the findings across different models. A baseline model predicted course grade using only gender/sex, high school GPA, and SAT math scores. Next, we added the mindset variables one-by-one in order of correlation strength with course grade until all mindset variables were included. All models with significant mindset variables were kept, along with the final model with all variables included as a robustness test. The regression analyses used an average across pre and post scores. Average scores were used instead of pre or post survey scores for two reasons. First, using post survey scores raises a question of causality (did course performance affect mindset or did mindset affect course performance?). Second, the average score is a proxy for students' mindset during the semester, while they were taking the course, rather than after the class. Using average rather than only pre survey data is particularly important given the sizable changes from pre to post alongside only moderate pre-post stability that were observed in several of the attitudinal variables. The results of linear regressions using either pre survey or post survey scores can be found in the supplemental material.

After using a linear model, we also implemented regression analysis using a threshold method based upon the low, medium, and high categories described in the preceding section. Instead of using continuous variables for each mindset score, these models used dummy variables for the two higher thresholds, treating low as the contrast group. For each mindset component, we performed a regression controlling for SAT math, High School GPA, and gender. Finally, if a mindset belief dummy variable predicted grade, we performed each regression again, but included an interaction between gender and that mindset belief dummy variable. Such an interaction test could reveal whether a dimension predicts course grade for one gender but not another or to a much larger extent for one gender.

7.3 Results

7.3.1 RQ1. Do physics intelligence mindsets organize into four factors (My Ability, My Growth, Others' Ability, and Others' Growth) as they did for students taking physics for engineering and physical science majors?

Initially, there were 19 items in the mindset survey, which can be found in Appendix H. A one-factor (in which all items were contained in the same construct) and both two-factor models (in which one model construct used "growth" and "ability" items, and another model used "me" and "others" factors) were rejected due to poor overall model fit. After deciding on a four-factor model, six additional survey items were removed (that is, they were completely excluded from the analysis) during the CFA model testing process due to low factor loadings or cross-loading that led to a poor model fit. The factor loadings for the 12 remaining items in the final model can be seen in Table 26. This model meets all chosen fit index cutoffs. Standardized factor loadings of each item were all greater than 0.5 [173, p.301], as seen in Table 26. All other fit indices (CFI, TLI, SRMR, and RMSEA) along with their cutoff values [223], can be seen in Table 27. Thus, answering RQ1, the final model had the four predicted mindset constructs: My Growth, In Table 26, Cronbach's α values were between 0.71 and 0.89 for all constructs for both the pre and post surveys, within the acceptable values range of 0.70 to 0.90 [108].

My Ability, Others' Growth, and Others' Ability. Further, each construct was based upon three items, similar to prior work on mindsets [16, 33, 65].

Finally, the upper left and lower right of Table 28 reveals that intercorrelations among the scales at pre and at post are all moderate and positive (after reverse coding of ability mindsets), but none are so high ($r > 0.7$) as to represent redundant measures. Correlations within each construct from pre to post (upper right of Table 28) showed low to moderate stability of the mindsets during this course experience, with especially low stability of the My Growth mindset.

7.3.2 RQ2. Gender Differences

7.3.2.1 a. Are there overall gender/sex differences in the means or distributions (in low, medium, and high categories) of students' physics intelligence mindset beliefs?

The winsorized means of both men's and women's mindset dimensions can be found in Table 29. As a reminder, all the mindset constructs are scored from 1 to 4, and are coded such that high scores correspond to a strong agreement with growth mindset beliefs or rejection of of a fixed mindset beliefs. Table 29 also shows the unpaired t-tests and Cohen's d effect sizes comparing men and women's mindsets at the beginning of the semester.

Answering the means component of RQ2a, as expected based upon prior work, men generally had higher mindset scores than women, as shown in Table 29. The smallest gender differences were non-significant, while the largest differences had moderate effect sizes. Men had higher mindset scores than women in every mindset category both pre and post.

Next, we divided students into groups that reported low (< 2.5), medium ($2.5 - 3.5$), and high (≥ 3.5) on a 1-to-4 scale for each mindset dimension. These student distributions can be seen in Figure 8a for the pre survey and Figure 8a for the post survey. As an important context note, more students in this course had growth rather than fixed mindsets: average scores for both genders are closer to 4 than to 1 for the pre survey. Distributions of mindset scores by gender showed similar trends to the means. For all constructs, men were more likely than women to fall into the high category, which can be seen in Figure 8. Similarly, for all constructs women were more likely than men to fall into the low category. There was only one exception to this trend, but this category did not have any student in the low category. Figure 8 also reveals that at the end of the semester men continue to be more likely to fall into the high category and are less likely to fall into the low category.

7.3.2.2 b. Are gender/sex differences in the means or distributions of students' physics intelligence mindset beliefs especially localized to particular dimensions?

Though men generally had higher mindset scores than women, the size of the gender differences varied by mindset construct, which we will now discuss individually. My Growth beliefs had no statistically significant mean gender difference between men and women at the start of the semester (see Table 29. For this construct, approximately half of women report high scores for pre beliefs (compared to 58% of men). By contrast, pre My Ability beliefs had the largest mean gender difference of any of the four pre constructs. These gender differences are also apparent in the score distributions that are found in Figure 8a. At pre, over half of men report a high My Ability score, while only one-third of women do.

The dimension that had the smallest pre gender/sex mean difference was Others' Growth. In this category, men and women had indistinguishable scores, shown in Table 29. Others' Growth also had the smallest gender differences in pre score distributions. Figure 8a shows that less than 5% of both men and women reported low pre scores, and a similar portion of men and women reported high scores (42% versus 43%).

Others' Ability had the lowest pre scores of any construct. Additionally, Others' Ability had statistically significant gender differences in pre scores. The Others' Ability gender differences are moderate at the start of the semester. As Others' Ability had the lowest scores of any construct, it also had the largest portion of students reporting low beliefs. In particular, 17% of women reported low pre Others' Ability beliefs, compared to $< 5\%$ of men. Men were also more likely than women to report high pre Others' Ability scores.

Broadly, we found that men tended to report higher mean pre scores in all four mindset constructs than women. The categorical distributions had a similar trend: women were more likely to report low scores and men were more likely to report high scores for all four pre mindset factors. However, to answer RQ2b, the gender differences were smallest in the My Growth and Others' Growth factors, and largest in the My Ability and Others' Ability factors.

7.3.2.3 c. Do gender/six differences grow or decline during students' first university-level physics course?

All students had statistically significant drops in mindset beliefs over time. The decreases in mean scores for each factor can be seen in Table 29 and the changes in student score distributions can be found in Figure 8. Men had similar but usually smaller decreases than women, with quantitative variations by construct.

The My Growth construct did not have statistically significant mean gender differences at the start of the semester. However, by the end of the semester, there is a moderate and significant mean gender difference in this construct. This growth in gender difference is the largest of any construct, and appears to be due to a dramatic drop ($d = -0.92$) in women's My Growth beliefs. This decrease is also shown in Figures 8a and 8b, which shows the distributions of students low, medium, and high mindset beliefs at post. At the end of the semester, only 19% of women have high My Growth beliefs, compared to 27% of men. Both men and women have lower My Growth beliefs at the end of the semester, but the shift was more drastic for women.

Turning to the mindset with the largest gender difference, Table 29 reveals that My Ability gender differences grew from $d = -0.51$ at pre to $d = -0.61$ at post. These gender differences are also apparent in the score distributions that are found in Figure 8. Both men and women show decreases in the number of students reporting high scores from pre to post, but at the end of the semester men were over twice as likely as women to report high My Ability scores. Women were also more than twice as likely as men to report low post My Ability scores.

By contrast, focusing on the dimension that had the smallest gender/sex differences, Others' Growth, men and women were indistinguishable at pre, and then there was a small gender difference at post. For pre Others' Growth, Figure 8 shows that less than 5% of both men and women reported low scores, and a similar portion of men and women reported high scores (42% versus 43%). Post Others' Growth distributions show minimal gender differences for low scores, but 7% more men than women report high My Growth scores at the end of the semester.

The construct with the lowest scores, Others' Ability, also showed the lowest (but still substantial) declines, and these declines were similar for men and women. Thus, the gender difference was similar at pre and post, but in the context of overall relatively low scores. At the end of semester, one-third of women and one-fifth of men had low scores in this construct. Women were also less likely than men to report high post scores.

In sum, to answer RQ3c, mindset scores generally declined over time, but they tended to decrease more for women than men, leading to larger gender differences at the end of the semester than at the start. The largest increases in gender differences from the start to the end of the semester were for My Growth and Others' Growth beliefs. The smallest increases in gender differences from the start to the end of the semester were for My Ability and Others' Ability beliefs. However, My Ability had the largest differences both pre and post.

7.3.3 RQ3. Do any of the mindset dimensions predict course grade and is the predictive relationship linear?

First, we conducted multiple regression analysis to find which of the four mindset beliefs best predicted physics course grade. We conducted this analysis two ways. First, we used linear regression using students' mean scores for each construct (see Table 30). Each model was conducted using the average of pre and post survey mindset scores due to the large changes in mindsets across the semester. Similar models using pre and post survey results can be found in the supplementary materials.

These models needed to include controls because there were also gender/sex differences in prior academic performance, although in opposing directions: for our sample women tended to have higher high school GPAs than men, but lower SAT math scores, as seen in Table 29. However, men in the sample tended to have higher physics 1 course grades than women. Here we investigate whether mindset differences could account for this gender difference in grade outcome.

Model 1, which can be seen in Table 30 uses only gender, SAT/ACT math scores, and HS GPA to predict students' physics 1 course grades. All three predictors are statistically significant. This model establishes that women had lower physics 1 course grades than men, even when controlling for High School GPA and standardized test scores, formally establishing that other factors are needed to account for gender/sex differences in course performance.

Model 2 includes My Ability as a fourth predictor. My Ability was chosen the first predictor to add because it has the strongest correlation with course grade for both the pre and post mindset components (see Table 28). Model 2 in Table 30 reveals that adding average My Ability to the model weakens the relationship between gender/sex and physics 1 grade, though it remains a statistically significant predictor. Model 2 also has an increase in Adjusted R-squared compared to Model 1. This means that Model 2 explains more of the variance in course grades than Model 1 even with a penalty for having an additional predictor [108].

Model 3 includes all four mindset components. This model reveals that both My Growth and My Ability are positively correlated with physics 1 course grades. Adding these other constructs marginally decreases the regression (β) coefficients of gender, SAT/ACT Math, and My Ability. Additional model testing revealed that My Growth average is a not statistically significant predictor of course grade unless either or both Others' Ability or Others' Growth are added to the model. Most importantly, Model 3 shows that the "My" dimensions of mindset (especially My Ability) are stronger predictors of physics 1 grades than "Others" dimensions.

Next, we conducted the regression analyses predicting grades using two dummy variables for each mindset construct: one for students categorized as reporting having medium mindsets and one for students categorized as having high mindsets; the regressions then treat the low group as the reference category [108, p. 551]. As a reminder, on the 4-point Likert scale, low was < 2.5, medium was $2.5 - 3.5$, and high was ≥ 3.5.

We focus on the models of this type that added dummy codes for each mindset factor individually, rather than using all four factors simultaneously. Models that used all mindset constructs simultaneously can be found in the supplementary materials. For each construct, we first introduce a model that predicts physics 1 course grade using each of the four mindset constructs (My Ability, My Growth, Others' Ability, and Others' Growth), gender/sex, SAT/ACT Math scores, and high school GPA. These models can all be seen in Appendix I. When both gender and mindset predictors were statistically significant, we proceeded to include interaction terms to test whether men and women's mindset scores predicted course grade to a different extent.

Looking across the factors, the dummy code approach replicated the high-level findings of the linear modeling approach in that strong agreement with My Growth beliefs was the best predictor of course grade. On the other hand, both medium and high agreement with My Ability beliefs predicted course grade (Figure 9). Other's Ability also was a statistically significant predictor, in contrast to being not significant as a linear predictor. On a related point, the support for linearity of effects was weak. Saliently, medium and high effects for Others' Ability were almost identical, potentially explaining why the linear model was not significant. However, even in the cases of My Ability and My Growth, the effect of high levels of agreement was not statistically different from medium levels. Others' Growth was not included in Figure 9 because neither medium nor high levels predicted course grade. Additionally, no statistically significant gender interaction term was found for any of the average mindset components. In other words, the relationship between mindset and course grade was similar for men and women, validating the use of simpler models that did not include interaction terms.

7.4 Discussion

Regarding RQ1, we found four components to students' mindsets (My Ability, My Growth, Others' Ability, and Others' Growth) using a survey instrument designed to specifically test for these components. This result replicated the findings of past work using previous iterations of the survey [16, 33]. These findings build on past work that separated mindset into multiple constructs, either between my versus others' mindset dimensions [201] or between growth and ability dimensions [192–194].

The four components were only moderately correlated with one another (18–34% shared variance at pre) and were separable in CFA models. Further, though each component showed similar patterns of gender difference, and change over time, the magnitudes of these effects were different and each component predicted course grades with different strengths. Thus, our components were not only separated by psychometric analyses, but by empirical patterns as well. Therefore, future research should avoid collapse measurement of mindsets into overall intelligence mindset scores.

Regarding RQ2, there were gender differences in the pre survey means of My Ability and Others' Ability, and men tended to report higher scores than women. There were gender/sex differences in the distributions of all mindset beliefs. More men fell into the high score range for each mindset component, and more women fell into the low score ranges. These differences were more pronounced for My Ability and Others' Ability pre. Women in this context were more likely than men to believe that physics requires innate ability. They were also more likely than men to believe that they did not personally have this innate ability. This is particularly concerning for this student sample, which consists of relatively high-achieving students who had decided to pursue bio- and health-science majors.

Even though women were the numerical majority in this context, the gender/sex differences in the means of each mindset component increased substantially from pre to post for My Growth and Others' Growth. As a result, men reported much higher mean scores than women for all four mindset constructs at the end of the semester. By the end of the semester, women were more likely to report low mindset scores than men in three out of four constructs (men and women reported low Others' Growth scores at similar rates). Men were more likely than women to report high mindset scores for all four mindset constructs. These inequities are not present because men had steady or increasing mindset beliefs. Instead, both men and women had moderate-to-large drops in all four mindset components during the semester. Women tended to have larger drops than men, creating or widening gender differences. It is important to note that, while there are gender inequities in student mindsets, the decreases over time are concerning themselves. Research in physics mindsets has found decreases in mindset beliefs for both men and women [16, 33]. Motivational factors, such as self-efficacy or interest, commonly tend to decrease over time in introductory physics coxfurses, both for physical science and engineering majors

[29,32], as well as for bioscience majors [25].

A large body of research shows that women tend to have lower motivational characteristics in physics courses than men [11,12,42], including mindsets [16,33]. Most of this research has been conducted in physics courses in which men outnumber women, such as courses for engineering and physical science majors. This study shows that gender differences also exist in courses in which women outnumber men, such as physics courses aimed at bioscience majors.

One similarity between mindset trends in these two types of courses is that students enter the course with gender differences in mindset. One important difference is that, while women in both categories of courses show moderate-to-large decreases in mindset scores, men show much larger decreases in mindset scores in courses aimed at bioscience, rather than engineering or physical science students [16,33]. In this context, efforts to increase mindset scores may benefit all students while simultaneously decreasing gender differences in physics intelligence mindset.

Regarding RQ3, we found that using linear regression models, both My Ability and Others' Ability positively predict physics 1 course grade. Mindset differences, especially related to My Ability, may offer a partial explanation for gender differences in physics 1 grades despite men and women having only small and opposing differences in SAT Math scores and HS GPAs. However, less than half the grade gender/sex difference was explained by the My Ability construct. The remainder of the gender/sex differences may be explained by other motivational or environmental factors [11,12,41,42,65,188].

Using threshold regression models (in which we divide students into groups with low, medium, and high scores for each construct), we found slightly different results. Here, we found that having at least a medium score for My Ability and Others' Ability positively predicted course grades. A high score for My Ability, My Growth, and Others' Ability all positively predicted course grades.

One theme that emerges in both the linear and threshold regression models is that agreement with both My Ability and My Growth beliefs positively predicts learning outcomes for both men and women. We also found that mindsets are malleable over time. If self-mindsets are more malleable and have a stronger correlation to learning outcomes, then one goal of physics instructors should be to focus on messages targeting the students' beliefs about themselves and their own growth rather than presenting students with messages about students or intelligence in general.

This study reveals that there are both similarities and differences between the mindsets of engineering/physical science majors and bioscience majors. Both groups had four mindset components: My Ability, My Growth, Others' Ability, and Others' Growth. Additionally, women in both groups saw declines in mindset component scores during the semester [16, 33]. Men in both groups saw significant declines in their My Ability and My Growth scores, but only men in the physics courses for bioscience majors saw significant declines in Others' Ability and Others' Growth [16, 33]. Finally, My Ability was the strongest predictor of course grades for both groups of students [16, 33]. However, for students taking physics for bioscience majors, My Growth and Others' Ability also predict course grades.

If an instructor wants to help students reject fixed mindsets and cultivate growth mindsets, student-level interventions can be valuable. In many such interventions, students are explicitly taught that hard work and effort, not innate ability, are necessary for success [132, 210]. These lessons are particularly important in physics, where fixed mindsets are commonplace [59].

Successful student-level mindset interventions tend to provide opportunities for self-reflection and show students that they can change their own intelligence. For example, students may be asked to remember instances during which they were able to improve their abilities [132, 210]. Interventions may also share stories from a diverse group of peers or experts about overcoming academic challenges, so that all students may find someone they can relate to. If a relatable role model shares that they were able to work hard to achieve success instead of relying on innate talent, students may realize they can do the same and develop growth mindsets for themselves [127, 132].

Instructor-focused change can be useful as well. Instructor mindsets can predict student achievement in their courses [211]. In addition, students in courses taught by instructors with growth mindsets report increased interest in their courses as well as fewer concerns about fair treatment and low grades [214]. Instructors with growth mindsets encourage students to accept mistakes and failures as a part of a normal learning process, congratulate persistence, praise effort rather than intelligence when students succeed [209, 212], and are more likely to implement active learning in their courses [213]. On the other hand, instructors with fixed mindsets tend to have low expectations of students they believe lack natural talent, which can lead instructors to give easier assignments or encourage students to drop difficult classes because of presumed low ability [212].

7.5 Conclusions

This study shows that intelligence mindset can be divided into four constructs: My Ability, My Growth, Others' Ability, and Others' Growth. Previous work in studying mindset has divided along either growth/ability or me/others categories, but rarely simultaneously. Next, this work reveals that gender/sex differences are more pronounced in the "Ability" categories than the "Growth" categories. Gender/sex differences developed or widened over the semester for all mindset constructs. These differences are the result of substantial drops in all four mindset factor scores from all students, which tended to be even larger for women than for men.

Next, students' mindset scores decrease over the semester for all four constructs. They also show women's mindset scores decrease more than men's. We also find that My Ability and My Growth consistently predict the course grade. My Ability positively predicts grades if students report a medium or high score, but My Growth only predicts grades if students report a high score. This information may be useful to target mindset interventions to student beliefs. A student who believes nobody can become more intelligent through hard work may have different needs than one who believes that most people can become more intelligent but that they personally lack the ability to do so.

7.6 Limitations and Future Directions

We now note several limitations of this study. First, the analyses were correlational in nature: the causal nature of physics intelligence mindsets would need to be further supported through intervention studies and interview data. The established benefits of other mindset interventions [209]) suggest such a causal link is plausible, and we note that future interventions that focus on My Ability and My Growth may show even larger effects.

Another limitation is the generalizability of the findings. The studied institution is predominantly white, so we were unable to study if mindset beliefs differ or predict grades differently for students of different racial/ethnic backgrounds due to low sample size. Because of the focus on gender/sex in this study, future work should also explicitly include students who fall outside of the binary gender/sex categories included here, as well as transgender students who may not have their gender accurately recorded by the university. Though this university recently began to include more sex/gender options for students, qualitative studies may be more appropriate to understand mindsets in these marginalized populations until student samples are large enough to be meaningful in quantitative analysis.

Table 26: Survey items included in the study and standardized factor loadings for pre and post surveys.

	Construct name or Item	λ	
		Pre	Post
	My Growth ($\alpha_{pre} = 0.83$, $\alpha_{post} = 0.89$)		
1.	I can become even better at solving physics problems through hard work	0.77	0.85
2.	I am capable of really understanding physics if I work hard	0.85	0.91
3.	I can change my intelligence in physics quite a lot by working hard	0.80	0.84
	My Ability ($\alpha_{pre} = 0.77$, $\alpha_{post} = 0.88$)		
4.	I won't get better at physics if I try harder	0.61	0.72
5.	I could never excel in physics because I do not have what it takes to be a physics person	0.80	0.88
6.	I could never become really good at physics even if I were to work hard because I don't have natural ability	0.84	0.90
	Others' Growth ($\alpha_{pre} = 0.85$, $\alpha_{post} = 0.81$)		
7.	People can change their intelligence in physics quite a lot by working hard	0.84	0.77
8.	If people were to spend a lot of time working on difficult physics problems, they could develop their intelligence in physics quite a bit	0.83	0.80
9.	People can become good at solving physics problems through hard work	0.74	0.80
	Others' Ability ($\alpha_{pre} = 0.71$, $\alpha_{post} = 0.75$)		
10.	Only a few specially qualified people are capable of really understanding physics	0.68	0.70
11.	To really excel in physics, people need to have a natural ability in physics	0.73	0.80
12.	If a student were to often make mistakes on physics assignments and exams, I would think that maybe they are just not smart enough to excel in physics	0.62	0.65

Table 27: Fit indices for the CFAs testing survey validity at pre and post, along with the applied fit index cutoffs. There were 419 students included in the factor analysis.

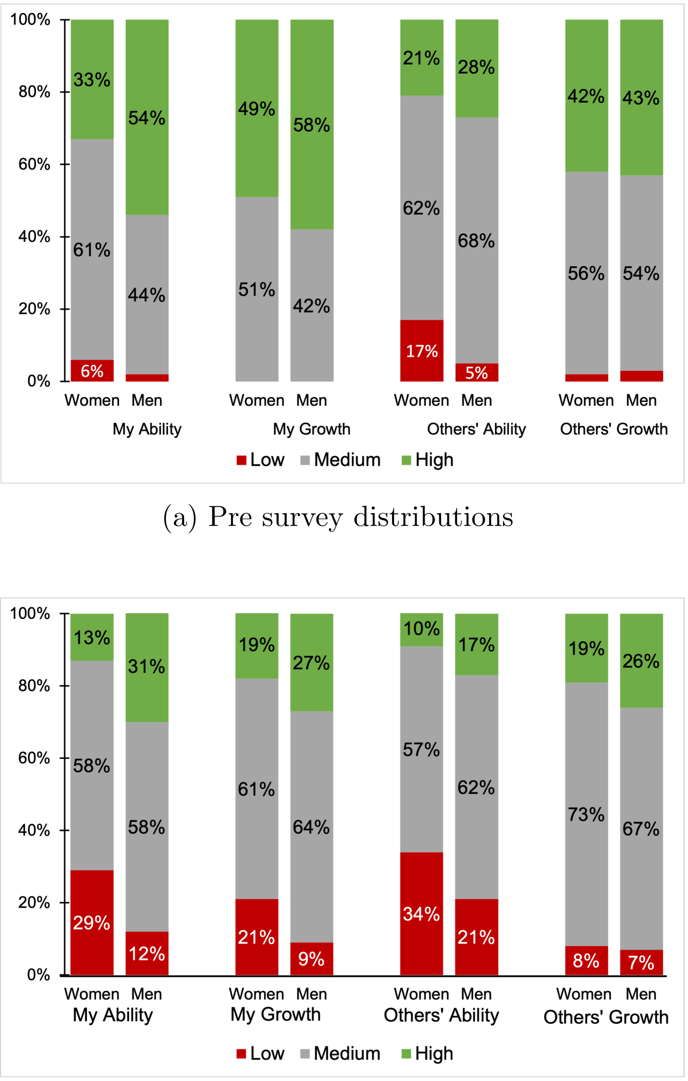
	CFI	TLI	RMSEA	SRMR
Cutoff	≥0.90	≥0.90	≤0.08	≤0.08
Pre	0.97	0.95	0.06	0.05
Post	0.97	0.95	0.07	0.04

Table 28: Pearson correlations between each mindset construct as well as physics 1 course grade. The following abbreviations are used: My Ability (MA), My Growth (MG), Others' Ability (OA), and Others' Growth (OG). $p < 0.001$ unless otherwise noted by: $^{*} = p < 0.05$, $^{**} = p < 0.01$, and ^{ns} = not statistically significant.

	Pre				Post				
	MG	MA	OG	OA	MG	MA	OG	OA	Grade
MG Pre					0.28				0.08^{ns}
MA Pre	0.53					0.47			0.10^{*}
OG Pre	0.58	0.47					0.36		-0.03^{ns}
OA Pre	0.42	0.51	0.43					0.44	0.06^{ns}
MG Post									0.25
MA Post					0.67				0.34
OG Post					0.59	0.48			0.16
OA Post					0.53	0.67	0.46		0.17

Table 29: Mean and standard deviation (SD) of each mindset factor, along with SAT Math, HS GPA, and Physics 1 grade, Cohen's d and t-test of gender/sex differences ("d Gender"). Positive values of d indicate that women had a higher score or that scores increase over time. $^{*} = p < 0.05$, $^{**} = p < 0.01$, and $^{***} = p < 0.001$.

		Women ($n = 276$)		Men ($n = 143$)		
Variable		Mean	SD	Mean	SD	d Gender
My Growth	Pre	3.45	0.47	3.54	0.45	-0.19
	Post	2.95	0.60	3.20	0.57	-0.42^{***}
	d Over Time	-0.92^{***}		-0.69^{***}		
My Ability	Pre	3.28	0.48	3.52	0.48	-0.51^{***}
	Post	2.82	0.64	3.20	0.57	-0.61^{***}
	d Over Time	-0.81^{***}		-0.63^{***}		
Others' Growth	Pre	3.37	0.47	3.40	0.48	-0.08
	Post	3.09	0.44	3.20	0.51	-0.24^{*}
	d Over Time	-0.62^{***}		-0.41^{***}		
Others' Ability	Pre	3.03	0.56	3.24	0.47	-0.40^{***}
	Post	2.73	0.60	3.00	0.61	-0.44^{***}
	d Over Time	-0.52^{***}		-0.47^{***}		
HS GPA		4.19	0.38	4.07	0.42	0.30^{**}
SAT/ACT Math		669	65	688	65	-0.29^{**}
Course Grade		2.91	0.76	3.22	0.71	-0.41^{***}



(a) Pre survey distributions

(b) Post survey distributions

Figure 8: Percentages of students who reported low (< 2.5), medium ($2.5 - 3.5$), or high (≥ 3.5) on a 4-point Likert Scale, by gender. Figure 8a contains pre survey distributions and Figure 8b contains post survey distributions. If any category contains $\leq 5\%$ of students, the percent is not labeled.

Table 30: Linear regression models predicting final course grade using average mindset beliefs. The regression (β) coefficiants are standardized, and the gender/sex variable was coded such that women = 1 and men = 0. $^{*} = p < 0.05$, $^{**} = p < 0.01$, and $^{***} = p < 0.001$. N=418.

Variable	Model 1	Model 2	Model 5
Gender	-0.18^{***}	-0.12^{**}	-0.11^{**}
HS GPA	0.34^{***}	0.32^{***}	0.32^{***}
SAT/ACT Math	0.39^{***}	0.39^{***}	0.38^{***}
My Ability Average		0.21^{***}	0.19^{**}
My Growth Avg			0.14^{**}
Others' Ability Avg			-0.04
Others' Growth Avg			-0.09
Adjusted R^2	0.37	0.41	0.41

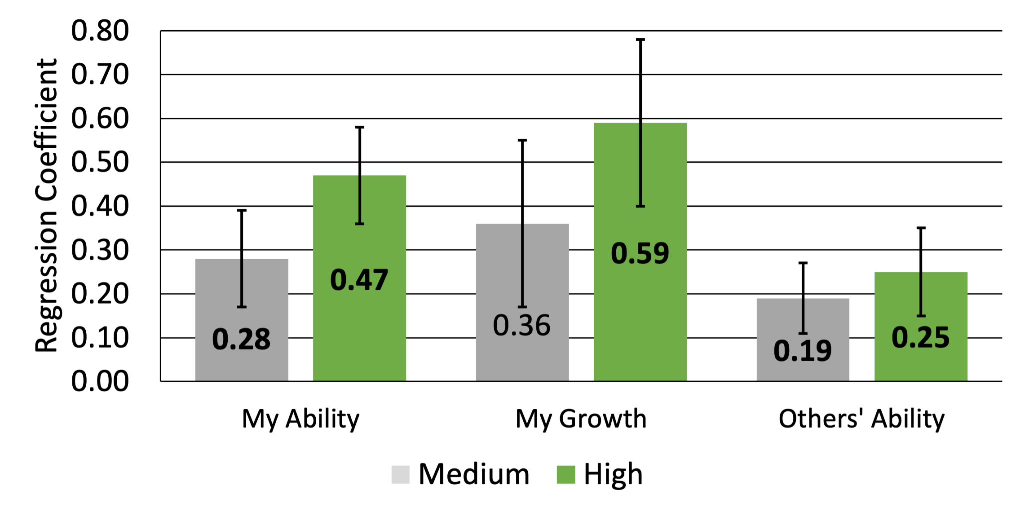


Figure 9: Unstandardized regression coefficients predicting course grade, controlling for gender, high school GPA, and SAT or ACT math scores. Error bars represent standard error. On a 4-point Likert scale, a medium score is $2.5 - 3.5$, and a high score is ≥ 3.5. Statistically significant regression coefficients are in bold. Others' Growth is excluded because it did not predict course grade.

8.0 Gender differences in grades versus grade penalties: Are grade anomalies more detrimental for female physics majors?

8.1 Introduction

In recent years, physics education researchers have been particularly focused on creating equitable learning environments, which is particularly important for underrepresented students such as women [5, 11, 41, 65, 158, 224–226], racial and ethnic minority students [158, 226–228], and students with disabilities [229, 230]. Here, we focus on women majoring in physics because they are more drastically underrepresented than women in many other science, technology, engineering, and mathematics (STEM) disciplines. Prior research has explored gender differences in performance and persistence in physics and other STEM fields [81, 82]. This research has explored a range of potential factors that may lead to such differences. Some factors that researchers have investigated include societal biases regarding who can be successful in physics [55, 59], lack of encouragement from families and instructors [67, 97, 208], and motivational characteristics and attitudes towards physics learning [11, 41, 224].

Societal stereotypes about physics are still prevalent, and both practitioners and students often believe that success in physics requires natural ability. For example, physics researchers are more likely than those in other STEM disciplines to endorse that their subject required natural ability for success [59], and young people often hold beliefs that "physics has always been seen as ... really hard, and you know you have to be so clever to understand it" [231]. This type of belief that physics requires a particular ability may combine with common societal stereotypes that men and boys are more likely than women and girls to be extremely intelligent [57, 222, 232], to discourage women from believing they have what it takes to pursue physics.

Lack of encouragement from families may also discourage early interest in science and physics. For example, parents tend to rate boys' abilities in math higher than girls' [67, 97], and are less likely to explain science to girls than boys while using interactive science exhibits at museums [181]. Because of these lack of early experiences and encouragement, girls may be less likely to develop an early interest in science. Similar lack of encouragement from instructors later in life may discourage women from participating in physics as well. For example, in one study [208] faculty members in physics rated men as more competent than women with an identical curriculum vitae. If women are not receiving the same amount of confidence or encouragement from their instructors, they may be less compelled to pursue physics.

Finally, some research on gender differences in physics performance and persistence has focused on motivational beliefs and attitudes towards physics learning [11, 41, 224]. One such attitudinal construct is academic self-concept, which describes a long-term expectation of success that students hold regarding their academic abilities and that depends on outside feedback, such as grades [79, 80, 233]. Low academic self-concept may lead to lower future achievement and persistence because it discourages student engagement in a domain [79]. When women leave STEM disciplines, particularly physics, they often do so with higher grades than the men who remain in the program [39, 53, 54]. For example, a recent investigation shows that among students with the same STEM GPAs, women were more likely to leave the major, while men were more likely to earn a degree [39].

There are many potential partial explanations that have been suggested regarding why women who are meeting or exceeding the requirements of their programs leave. These include lack of role models, societal stereotypes and biases about who can excel in these disciplines [38, 56, 57, 59, 67, 202], gender discrimination in hiring [208], and differences in STEM motivational beliefs such as self-efficacy [6–12, 23, 32, 34, 41, 42, 65, 188], recognition from instructors and peers [27, 29], intelligence mindset [16, 33], and sense of belonging [22, 24]. One related reason for why this happens may be lower academic self-concept of female students in these courses compared to male students. Though none of these factors may provide complete explanations of gender differences in physics, aiming to address them simultaneously may create a better learning environment for women in these programs.

Here, we focus on physics majors and inquire about gender differences in grade penalties. In order to quantify grade penalty, we define grade anomaly as the difference between a student's grade in a course of interest and their grade point average (GPA) in all other classes up to that point. The mean of this statistic for all students who took a course is the average grade anomaly (AGA). We divide average grade anomalies into "bonuses" and "penalties". A course in which students on average earn a lower grade than usual has an AGA with grade penalty, while a course in which students on average earn a higher grade than usual has an AGA with grade bonus.

Within our framework, we posit that grade anomaly may allow us to track, through institutional grade data, an important measure of how courses may affect students' academic self-concept. Our framework uses grade penalty as a central construct instead of grade because students' academic self-concept is often based on comparisons, not absolute grades [78]. Students may compare their grades across courses to determine which disciplines they excel at or struggle with [78]. Additionally, students tend to have a fairly fixed view of what "kind" of student they are, e.g., students may endorse the idea that "If I get As, I must be an A kind of person. If I get a C, I am a C kind of person" [53]. Grade anomalies may challenge or reinforce students' ideas about what kind of student they are, and if they are capable of succeeding in their chosen major. Many students who leave STEM majors explicitly cite lower grades than they are used to as a reason for doing so [53, 54]. Grade penalties are more common and extreme in STEM disciplines than in humanities or social science departments [53, 152, 234, 235], and women tend to have larger grade penalties than men in many subjects, including physics [152].

In this paper, we use Situated Expectancy Value Theory (SEVT), studies about why students leave STEM, and previous work on grade anomalies to explore whether the average grade anomalies for male and female physics majors are different, making grade anomalies an equity issue in physics. We also posit that grade anomaly may be a better measure of self-concept [78] than raw grades because it is a unique measure of "within-student" frame of reference (i.e., students are comparing their own grades across different courses as opposed to comparing their grades with others) [233].

8.1.1 Research Questions

We aim to answer the following research questions regarding average grade anomalies (AGAs):

RQ1. For which of their courses do students majoring in physics, on average, receive a "grade penalty" and for which courses do they receive a "grade bonus"?

RQ2. To what extent do men and women have different AGAs in their physics courses?

RQ3. To what extent do gender differences in AGAs follow the same trends as gender differences in average grades?

8.1.2 Theoretical Framework

Expectancy Value Theory (EVT) [79] and Situated Expectancy Value Theory (SEVT) [78] are frameworks to understand student achievement, persistence, and choice of tasks in a domain (e.g., physics or chemistry). EVT posits that performance and persistence is determined by someone's expectation of success and the extent to which they value that task. If a student expects they will succeed in a task and believes that the task will be valuable to them (for personal interest, as a path to achieve another goal, etc.) they are more likely to pursue that task. If they do not expect to succeed and do not value a task, they are unlikely to attempt it. Here, we will focus primarily on student expectancies, though value is also important to understanding why some students may persist while others do not. Expectancies are a combination of academic self-concept, expectations for success, and perceptions of task difficulty [78–80, 233]. Academic self-concept is the most stable and the least task and domain-dependent of the three, and it is based primarily on grades and outside (e.g., from parents, peers, and instructors) feedback [78–80,233,236]. Grades can inform academic self-concept as both an external ("How good at math am I compared to other students?") and internal ("How good am I at math compared to English?") frame of reference [79, 80, 233].

Expectancies of success are domain and task specific, and refer to a student's belief in their ability to complete a specific task; which will include considerations such as knowledge and skill related to the subject, time allotted, and experience in a subject [78–80, 233]. Expectancy for success closely relates to Bandura's theory of self-efficacy [78, 80, 127]. A student may have a positive academic self-concept in math, but may have low expectancy for success if they take a math test on very new material they have not had adequate time to learn. The third issue related to expectancy, perceptions of task difficulty, is more straightforward; most students have less faith in their ability to do well on an exam if their peers have reported it to be particularly difficult [79].

In EVT, the three expectancy concepts were collapsed into one factor. However, the updated framework, SEVT, has called for a separation of these three concepts [78]. According to Eccles and Wigfield [78], combining academic self-concept, expectancies for success, and perceived task difficulty has led to a lack of understanding of the unique developmental mechanisms of each and how the three concepts relate. We posit that grade anomaly may be a better measure of how students' self-concept evolves [78] due to feedback about performance than raw grades. This is because students often judge their ability by comparing their grades across courses rather than comparing their grades to other students' (in EVT/SVET, this is called the "within student" frame of reference). Poor performance from a within-student frame of reference may cause students to question if they should continue in a discipline [78].

Grade anomalies allow us to measure how courses can affect students' academic self-concept [78, 80, 234] using institutional grade data [152], which may be more accessible to instructors and researchers than surveys or interview data. While students' raw grades are a useful measure, e.g., because they allow for direct comparison between students and because they are used by institutions to award scholarships and track student academic standing, we propose that using grade penalty in addition to raw grades gives researchers and instructors more insight into student self-concept (that is, a students' view of what "kind" of student they are).

A student who receives lower grades in their science courses than their humanities courses may take this as a sign that they are not capable of excelling in the sciences, even if the grades they earn are high enough for them to continue in their major

[53, 54]. This experience may be common, because grade penalties tend to be more extreme and widespread in STEM disciplines than in other subjects [53,235,237,238]. Women may also have fewer resources available to cope with these grade penalties than men are, in part due to lack of role models and societal stereotypes about who belongs in these disciplines and can excel in them [1, 55, 237]. When students leave STEM fields, they often list lower grades than expected as a reason [53, 54], and women tend to leave these fields with higher grades than men [39]. Larger grade penalties may be one potential reason for this discrepancy.

Several studies [152, 235, 239] have utilized "grade anomaly" or "grade penalty", the difference between a students' GPAs excluding a course of interest and their grades in all courses thus far. Koester et al. [235] conducted the first study we know of that focuses on average grade anomaly (AGA). They used AGA because it was perceived to be a better measure of how students view their comparative performance than their raw grades across different courses. They found that, at their institution, grade penalties were greater for STEM than non-STEM courses. Further, within STEM courses, grade penalties were smaller for men than women. In particular, they found that physics courses had the largest grade penalty and largest gender difference in AGA. The researchers theorized that large grade penalties and gender differences may be partially attributed to high-stakes assessments [13,94,120,121,240] and stereotype threat [77]. The Matz et al. [152] study had similar findings but with a larger student sample across multiple institutions. Across five universities, STEM courses had larger grade penalties and larger gender differences in AGA that usually favored men. Their study also raised concerns over high-stakes assessments, . They emphasized that large grade anomalies often reflect grading decisions made by instructors (for example, choosing high-stake assessments may increase gendered grade differences [13,94,119,120]), rather than being an accurate measure of student learning.

Additionally, Witteveen and Attewell [239] found that having lower STEM GPAs than overall GPAs during the first two semesters of university were negatively correlated with completing a STEM degree, even when controlling for gender, race, high school preparation, and college performance. This was not the case for courses taken later in students' college career, which speaks to the importance of introductory courses in student retention. Past research has found that during times of transition, the usually-stable academic self-concept becomes more dependent on grade feedback and less dependent on outsider (e.g., parental) feedback [233]. These findings hint at the importance of monitoring and reducing grade penalties in students' first few semesters.

Thus, past work provides evidence for the existence of average grade penalties in many STEM courses, and the existence of gender differences in these anomalies. Here, we present an investigation that focuses on average grade anomaly in various courses for physics majors. We analyze data to study if these trends hold in a more homogeneous population of students in the same major at a large university in the US, rather than combining students across institutions and many majors. We hypothesize that grade anomalies, e.g., grade penalty, discussed here can negatively influence students' internal frame of reference.

8.2 Methodology

8.2.1 Participants and Procedures

Participants in this study were students who declared a physics major at any time during the thirteen-year duration of the study. All participants attended the University of Pittsburgh, which is a large, public, and urban institution. Thus, our sample included both students who graduated with a physics major and students who declared a physics major but later switched to another field. We excluded summer introductory courses they are not a typical representation of courses at our institution. For example, many summer students do not primarily attend our institution, but are local students visiting home for the summer. In addition, summer courses are typically taught by graduate students, the class sizes are an order of magnitude smaller than those in the Fall and Spring semesters, many students work full time while taking an introductory course which is not as common during the Fall and Spring semesters. We had a total of 671 students who took 23,154 classes (including 5,713 physics courses). The sample consisted of 23.2% women and 76.8% men. All students in the sample selected one of these binary gender options. Students identified with the following races/ethnicities: 79% White, 9% Asian, 3% Hispanic/Latinx, 4% multiracial, 2% African American/Black, and 3% unknown or unspecified. This research was carried out in accordance with the principles outlined in the University of Pittsburgh Institutional Review Board (IRB) ethical policy, and de-identified demographic data were provided through university records.

We chose the twenty most common physics courses taken by students in our sample, many of which are mandatory for physics majors. All the courses we studied are listed in Table 31, along with information about the year in which the students typically take the course. The total number of students who took each course, as well as the gender distribution for each course can be found in Table 33. Due to changing course requirements over time (for example, Quantum Mechanics 1 was optional for the first three years of the study) and the tendency of students to leave the physics major over time [45, 53, 54, 156], more students took introductory than advanced courses.

8.2.2 Measures

8.2.2.1 Course Grade

Course grades were based on the 0-4 scale used at our university, with A = 4, B = 3, C = 2, D = 1, F = 0 or W (late withdrawal), where the suffixes '+' and '-', respectively, add or subtract 0.25 grade points (e.g., B- = 2.75 and B+ = 3.25), except for the A+, which is reported as 4. We are unable to report detailed grading schemes of each physics instructor, type of course, or any other detailed course-level information but a majority of courses are traditionally taught primarily using lectures.

8.2.2.2 Grade Anomaly

Grade anomaly (GA) was found by first finding each student's grade point average excluding the course of interest (GPA_{exc}), including all courses taken prior and simultaneously with the course of interest. This was done by using the equation

$$GPA_{exc} = \frac{(GPA_c \times Units_c) - (Grade \times Units)}{Units_c - Units} \tag{3}$$

where GPA_c is the student's cumulative GPA, $Units_c$ is the cumulative number of units the student has taken, $Grade$ is the grade the student received in an individual course, and $Units$ is the number of units associated with an individual course. After finding GPA_{exc} we can calculate GA by finding the difference between a student's GPA_{exc} and the grade received in that class:

$$GA = Grade - GPA_{exc}. \tag{4}$$

A negative GA corresponds to a course grade lower than a student's GPA in other classes (a "grade penalty"). A positive GA corresponds to a course grade higher than a student's GPA in other classes (a "grade bonus"). Average grade anomaly (AGA) is the mean of students' grade anomalies (GA) for each course, and is the metric by which we compare courses.

8.3 Results

8.3.1 For which of their courses do physics students receive a "grade penalty" and for which courses do they receive a "grade bonus"?

To answer RQ1, we calculated average grade anomaly (AGA) for our courses of interest. We show the descriptive statistics for both grades and AGA in Table 39. On average, students received grade penalties in most of their courses. However, two courses tended to give students grade bonuses: Introductory Physics Lab and Modern Physics Lab. Three courses gave neither a grade penalty nor a grade bonus (i.e., the AGA of the course was within one standard error of zero) : Optics Writing, Honors Introductory Lab, and Introduction to Astronomy. The courses that gave students grade bonuses or no grade anomalies were all lab courses except Astronomy and the Optics wiring practicum.

Four courses gave students particularly large grade penalties (≤ -0.50): Modern Physics 1, Thermodynamics and Statistical Mechanics, Mechanics, and Electricity and Magnetism 1. This means that a student taking one of these courses can expect to receive half a letter grade lower in these courses than they do on average. Notably, the courses with the largest grade penalties are mandatory (see Table 39).

The courses in Table 32 do not average to an AGA of zero because student sample changes over time due to students who leave the major and because not all courses students take are included. For example, students may receive grade bonuses in general education courses which students can select from several hundred courses.

Women and men had statistically significantly different outcomes in their first year courses, as shown by the MANOVA analysis in Table 34. Figure 33 displays the mean and standard error of both men and women's AGAs in first year courses. From Figure 10, one can see that there is a distinguishable difference between men and women's AGAs for each of the first year courses, and that men tended to have smaller grade penalties than women. The largest gender differences shown in Figure 10 and Table 33) tend to be in the courses that students take earliest in their time as physics students: Physics 1 and Physics 1 Honors. Thus, it is not surprising that first-year courses show a statistically significant gender difference in Table 34.

However, Table 33 also shows that there are no statistically significant gender differences for courses taken primarily by second, third, and fourth-year students. This finding is supported by Figure 10, which shows that women had indistinguishable AGA outcomes to men in most individual courses, including Honors Introductory Physics Lab, Electronics Lab, Introduction to Astronomy, Modern Physics 1, Modern Physics 2, Modern Physics Lab, Thermodynamics and Statistical Mechanics, Optics, Optics Writing, Electricity and Magnetism 1, Electricity and Magnetism 2, Computational Methods, and Quantum Mechanics 1.

Though these courses are not included in any course groups that show significant AGA gender differences, we note that women had favorable AGA outcomes (e.g., smaller grade penalties or larger grade bonuses) compared to men in two courses: Introductory Physics Lab and Quantum Mechanics 2.

8.3.2 To what extent do gender differences in AGAs follow the same trends as gender differences in average grades?

Next, we explore if grade anomaly and raw grades can provide different information. That is, does calculating AGA reveal additional trends beyond what raw grades can provide? Table 34 shows that the only group of courses that has a statistically significant gendered AGA difference is courses taken primarily by first-year students. Similarly, Table 34 shows that the only group of courses that has a statistically significant gendered grade difference is also courses taken primarily by first-year students. However, it is important to note that the gender difference is larger for AGA than raw grades.

This trend is further supported by Figures 10 and 11. These figures show that the gendered grade differences appear to be larger (i.e., their standard errors are further from overlapping) for AGAs than for grades, especially for first year courses such as Physics 1 and Physics 1 Honors. This trend is also shown by the between-gender effect sizes listed in Table 33: Cohen's d [109] is larger ($\Delta d = 0.18$) for AGA than grade for Physics 1, which is most students' introduction to the major.

Beyond first-year courses, there are a few differences in effect size between some courses, which are shown in Table **??**. For example, both Honors Introductory Lab

($\Delta d = 0.20$) and Modern Physics Lab ($\Delta d = 0.15$) appear to strongly favor women in terms of raw grades, but this effect is smaller for AGAs. Similarly, Modern Physics Lab appears to have favorable outcomes for women in raw grades (seen in Figure 11), but not AGAs (seen in Figure 10). Importantly, there are some additional trends revealed in AGA that cannot be seen with grade data alone.

8.4 Discussion

Our results show that there are grade penalties in the majority of courses studied. First, we discuss why grade penalties can potentially be harmful. It is important to note that students at the University of Pittsburgh do not declare their major until the end of their second year, so we are unable to track students who decided to leave the physics discipline before their third year. However, lower than expected grades, even in one course, can be a catalyst for students to leave STEM majors [53, 54]. This does not just include D and F grades or withdrawal from the course, but grades that were high enough to continue in the program that did not meet a student's personal expectations [53, 54]. This can especially be an important issue among high-achieving students, who are more likely to endorse perfectionism and feeling that their identity as "good STEM students" is threatened by B's and C's, or even a low grade on a single exam [53]. Thus, we hypothesize that the courses with largest grade penalties, in this case Modern Physics 1, Thermodynamics and Statistical Mechanics, Mechanics, and Electricity and Magnetism 1, are the courses that are more likely to discourage most students from continuing in physics.

In addition to seeing evidence of grade penalties in some physics courses, we also see evidence of gender differences in average grade anomalies in over half of the courses studied, particularly Physics 1, Physics 2, and Physics 1 Honors. We find these introductory courses to be particularly concerning. Our research shows that women have statistically significantly lower grades and AGAs than men in first year courses (see Table 33), which has the potential to affect women's academic self-concept more than other courses and they are taken during students' first year at the university when their academic self-concept is in major flux [233]. Additionally, prior research suggests that low STEM GPAs during students' first year are correlated with lower degree completion [239]. Because women leave majors with higher grades than men who remain both in [39, 53] and outside [237] of STEM, this raises serious equity concerns. Past work also suggests that women tend to have lower self-efficacy and sense of belonging that relate to academic self-concept [12, 41, 94, 224, 241, 242]. Women report feeling more demoralized than men when they receive lower grades than expected, and cite more worry about not understanding the material even if they receive A's, B's, or C's (all of which are grades that allow students to continue in most programs) [53]. This trend can be particularly strong among high-achieving women [53].

We hypothesize that women may be more likely to have a low academic-self-concept than men at similar performance levels for several reasons. First, prior research suggests that women are less likely to separate their academic self-concept from their grades which is one of the clearest types of recognition in a domain [53, 54, 237]. In particular, grades are the resource that women have the most access to. Academic self-concept is formed through grades and feedback from outsiders. Because women are less likely to receive recognition as someone with potential in STEM from their parents [67, 97, 181], society at large [55, 57], and their instructors [56, 130, 208], they are more likely to rely on grade information to develop their academic self-concept. Also, women often tend to earn higher grades than men with the same standardized test scores [53, 203]. Because women are often more accustomed to higher grades, they may have more concern about grades that are lower than what they are accustomed to, or they may compare their relatively-low STEM grades and leave for another subject that gives them the recognition for their work that they are accustomed to [53, 54].

We find that average grade anomalies and raw average grade data do not always reveal the same trends. Some courses have larger gender differences in AGA than in grades, such as Physics 1 and 2, Physics 1 Honors, Modern Physics Lab, and Optics Writing. This further speaks to the usefulness of tracking both AGA and grades of the students. An instructor may see a small or negligible grade difference by gender and assume that there is gender equity in their classroom based upon grade outcomes by gender, but without knowing the gender differences in AGA, the instructor will not understand how those grades are perceived by female and male students. Understanding both grades and AGA differences may allow instructors to understand both classroom-level inequities and the extent to which their course may be pushing out students, particularly those from underrepresented groups, such as women, out of STEM fields.

From our results, we make several recommendations to instructors and departments. First, measuring grade anomaly in addition to grades may be a useful way to find inequities in the learning environment. Measuring grades and gendered grade differences is both valuable for and accessible to individual instructors, but grade anomalies may be useful to departments concerned about students' retention over longer periods and finding which courses may be discouraging students, particularly those from underrepresented groups, to leave a major.

8.5 Conclusion and Future Research

In this investigation, we found that grade anomalies exist for all studied physics courses at our institution for those who had declared a physics major. Further, several physics courses had an average grade anomaly that favored men over women. These findings raise particular concern about the need for an equitable learning environment and outcome for these students. These results are very important because they provide some evidence that courses in physics departments tend to have grade penalties. They support the grade penalties found in introductory courses from prior work on grade anomalies. The relatively new measure of AGA may also act as one measure of academic self-concept that is easy for institutions to access and evaluate over time. This can also be useful to researchers as they develop separate measurements for academic self-concept and expectancies for success. Although we have strong evidence of grade penalties in the studied physics courses for those who had declared a physics major as well as gendered AGA differences, we did not have access to syllabi or other information about individual courses offered over the thirteen-year period of data collection. Therefore, we are not able to pinpoint specific practices that may lead to grade penalties, grade bonuses, or gender inequities at our institution. However, we know that out of the courses currently offered, most of these courses focus on teaching in a traditional, lecture-based, and exam-reliant format.

Finally, this research is based at a primarily white, large, public university. While our results may generalize to similar institutions [152, 235], we do not know what patterns of grade anomalies exist minority-serving institutions or community colleges. Conducting research at a diverse range of institutions, as well as a focus on how grade anomaly affects students from a variety of underrepresented groups, will help us more fully understand how grade anomalies differ for a range of students.

Table 31: Names of the courses studied and the percentage of students who take each course in a given year.

Course	1^{st}	2^{ed}	3^{ed}	4^{th}	$\geq 5^{th}$
Physics 1	**74**	19	4	2	1
Physics 2	**54**	34	9	1	2
Physics 1 Honors	**87**	8	4	0	1
Physics 2 Honors	**83**	9	4	2	2
Introductory Laboratory	17	**58**	15	7	3
Intro to Astronomy	35	**41**	15	5	4
Hon Introductory Laboratory	4	**75**	13	6	2
Modern Physics 1	6	**49**	27	13	5
Modern Physics 2	1	**44**	23	24	8
Electronics Laboratory	2	**43**	32	14	9
Modern Physics Laboratory	0	32	**44**	16	8
Thermodynamics and Statistical Mechanics	2	10	**42**	32	14
Mechanics	4	30	**41**	17	8
Electricity and Magnetism1	2	14	**53**	22	9
Electricity and Magnetism 2	0	7	**42**	34	17
Computational Methods	2	13	**44**	25	16
Optics	0	4	29	**47**	20
Optics Writing	1	12	28	**42**	17
Quantum Mechanics 1	0	4	25	**50**	21
Quantum Mechanics 2	0	4	25	**49**	22

Table 32: Mean and standard deviation (SD) of average grade anomalies (AGA) and grades, number of students (N) for each course, and the percentage of women that took the course ("% W"). Courses are marked as required or optional. Optional courses are chosen from a group of electives to fulfill degree requirements. We used the following abbreviations: Laboratory (Lab), Astronomy (Astro), Modern Physics (Modern), Thermodynamics and Statistical Mechanics (Stat Mech), Electricity and Magnetism (E&M), Computational (Comp), and Mechanics (Mech). Students may take either Physics 1 and 2 or Physics 1 and 2 Honors, and either Introductory Lab or Honors Introductory Lab

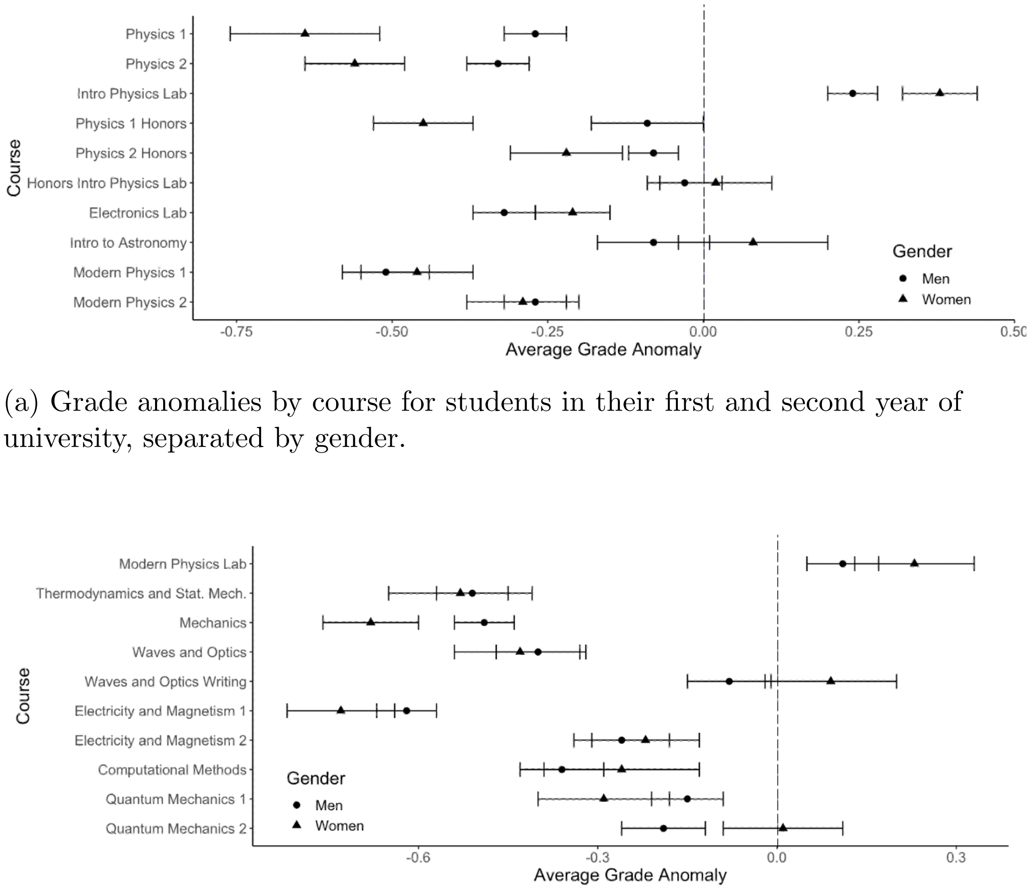
				AGA		Grade	
Course	Course Type	% W	N	Mean	SD	Mean	SD
Physics 1	Required	23	335	-0.36	0.92	2.89	0.92
Physics 2	Required	26	389	-0.39	0.86	2.87	0.94
Physics 1 Honors	Required	25	198	-0.18	1.06	3.31	0.77
Physics 2 Honors	Required	21	183	-0.11	0.54	3.43	0.70
Intro Lab	Required	23	371	0.27	0.71	3.49	0.82
Intro to Astro	Optional	23	151	-0.04	0.88	3.10	0.99
Honors Intro Lab	Required	15	116	-0.02	0.58	3.53	0.69
Modern 1	Required	19	315	-0.50	1.03	2.84	1.08
Modern 2	Optional	19	286	-0.28	0.71	3.04	0.93
Electronics Lab	Optional	22	241	-0.29	0.66	3.00	0.89
Modern Lab	Optional	19	125	0.13	0.29	3.46	0.63
Thermo	Required	21	219	-0.52	-0.78	2.88	1.05
Mechanics	Required	20	291	-0.53	0.70	2.83	0.93
E&M 1	Required	20	318	-0.64	0.83	2.68	1.08
E&M 2	Optional	23	100	-0.25	0.62	3.33	0.80
Comp Methods	Required	19	272	-0.34	1.00	2.95	1.24
Optics	Optional	20	196	-0.41	0.81	2.90	1.06
Optics Writing	Optional	25	166	-0.04	0.78	3.30	0.98
Quantum Mech 1	Required	19	204	-0.17	0.81	3.27	0.99
Quantum Mech 2	Optional	23	103	-0.15	0.59	3.41	0.74

Table 33: Average grade anomalies (AGAs), grades, and effect sizes for both measures for each course of interest, by student gender. Cohen's d is positive if men had higher grades than women in a course. We used the following abbreviations: Honors (Hon), Laboratory (Lab), Astronomy (Astro), Physics (Phys), Electronics (Elect), Thermodynamics and Statistical Mechanics (Stat Mech), Electricity and Magnetism (E&M), Computational (Comp), and Quantum Mechanics (QM).

	Women				Men					
	AGA		Grade		AGA		Grade		Cohen's d	
Course	Mean	SD	Mean	SD	Mean	SD	Mean	SD	AGA	Grade
Physics 1	-0.64	1.03	2.73	0.93	-0.36	0.92	2.94	0.92	0.41	0.23
Physics 2	-0.56	0.79	2.73	0.96	-0.33	0.88	2.92	0.93	0.27	0.20
Phys 1 Hon	-0.45	0.59	3.09	0.71	-0.09	1.16	3.38	0.78	0.34	0.39
Phys 2 Hon	-0.22	0.54	3.29	0.68	-0.08	0.54	3.47	0.70	0.25	0.26
Int Lab	0.38	0.51	3.62	0.58	0.24	0.76	3.45	0.88	-0.19	-0.20
Int to Astro	0.08	0.69	3.19	0.94	-0.08	0.92	3.07	1.01	-0.18	-0.12
Hon Int Lab	0.02	0.39	3.69	0.49	-0.03	0.60	3.50	0.71	-0.08	-0.28
Mod Phys 1	-0.46	0.72	2.86	0.91	-0.51	1.09	2.84	1.12	0.04	-0.02
Mod Phys 2	-0.29	0.64	3.06	0.78	-0.27	0.72	3.04	0.96	0.02	-0.02
Elect Lab	-0.21	0.43	3.10	0.73	-0.32	0.71	2.97	0.93	-0.15	-0.14
Modern Lab	0.23	0.50	3.64	0.51	0.11	0.60	3.42	0.65	-0.20	-0.35
Stat Mech	-0.53	0.79	2.97	0.96	-0.51	0.78	2.86	1.07	0.01	-0.10
Mechanics	-0.68	0.66	2.69	0.85	-0.49	0.71	2.88	0.95	0.28	0.20
E&M 1	-0.73	0.70	2.64	0.93	-0.62	0.86	2.69	1.11	0.13	0.05
E&M 2	-0.22	0.44	3.33	0.61	-0.26	0.67	3.33	0.86	-0.07	0.00
Comp Meth	-0.26	0.92	3.10	1.08	-0.36	1.02	2.91	1.28	-0.10	-0.15
Optics	-0.43	0.70	2.94	1.01	-0.40	0.84	2.89	1.07	0.03	-0.05
Optics Wri	0.09	0.70	3.48	0.83	-0.08	0.80	3.23	1.03	-0.22	-0.25
QM 1	-0.29	0.70	3.17	0.91	-0.15	0.83	3.29	1.01	0.17	0.13
QM 2	0.01	0.49	3.58	0.65	-0.19	0.61	3.35	0.76	-0.35	-0.31

Table 34: Three multivariate analyses of variance (MANOVA) are reported, with courses grouped to reduce listwise deletion into courses typically taken in students' first, second, third, and fourth years in the physics program.

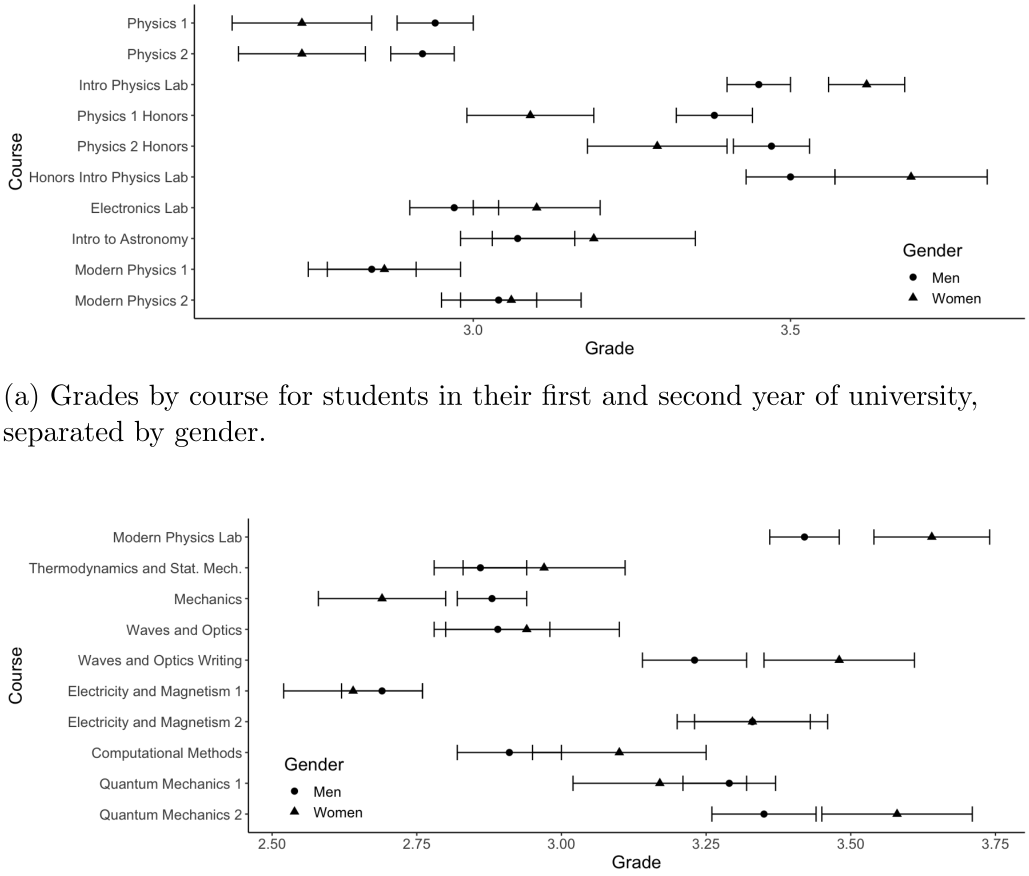
	Courses	AGA	Grade
1	Physics 1, Physics 2, Physics 1 Honors, Physics 2 Honors	F(1,1112) = 18.61***	F(1,1118) = 13.99***
2	Intro Lab, Intro to Astro, Honors Intro Lab, Lodern Physics 1, Modern Physics 2, Electronics Lab	$F(1, 1416)$ =2.22	$F(1, 1416)$ =3.23
3	Modern Lab, Stat Mech, Mechanics, E&M 1, E&M 2, Comp Methods	$F(1, 1383)$ =0.04	$F(1, 1383)$ =0.07
4	Optics, Optics Writing, QM 1, QM 2	$F(1, 702)$ =1.20,	$F(1, 702)$ =1.72



(a) Grade anomalies by course for students in their first and second year of university, separated by gender.

(b) Grade anomalies by course for students in their first and second year of university, separated by gender.

Figure 10: Average grade anomalies (AGA) of all students who declared a physics major by course, separated by gender. Ranges represent standard error of the mean. All courses are required except for Intro to Astronomy, Modern Physics 2, Electronics Lab, Modern Physics Lab, Electricity and Magnetism 2, Waves and Optics, Waves and Optics Writing, and Quantum Mechanics 2. The dashed line represents an average grade anomaly of 0.



(a) Grades by course for students in their first and second year of university, separated by gender.

(b) Grades by course for students in their third year of university and beyond, separated by gender.

Figure 11: Average grades of all students who declared a physics major by course, separated by gender. Ranges represent standard error of the mean. All courses are required except for Intro to Astronomy, Modern Physics 2, Electronics Lab, Modern Physics Lab, Electricity and Magnetism 2, Waves and Optics, Waves and Optics Writing, and Quantum Mechanics 2.

9.0 Impact of Grade Penalty in First-Year Foundational Science Courses on Female Engineering Majors

9.1 Introduction and Theoretical Framework

Gender differences in performance and persistence in science, technology, engineering, and mathematics (STEM) are well-studied phenomena, especially in engineering [1, 4, 44, 115]. When women leave STEM disciplines, they often do so with higher grades than men who remain in the program [53, 54, 180]. Women are more drastically underrepresented in engineering than many other STEM disciplines [1, 4, 115], so focusing on retention is particularly important for this field. If women are leaving engineering programs with grades that meet or exceed minimum requirements [53,54], it is likely that many students who would succeed in engineering careers will pursue other professional paths. There are many partial explanations regarding why women who are meeting or exceeding the requirements of their programs leave. These include societal stereotypes and biases about who can excel in these disciplines that discourage women from pursuing STEM careers [9,38,56,57,59,67,202,243,244], gender discrimination in hiring [208], and differences in STEM motivational beliefs such as self-efficacy and sense of belonging [6–8, 11, 12, 41, 44, 65, 115, 188]. We have been focusing on how to improve equity and inclusion in STEM, with a particular focus on motivational factors [9, 10, 22, 23, 32, 171].

Here, we focus on first-year engineering majors and introduce a framework that posits that grade penalty in first year foundational science courses for engineering majors may be particularly damaging to female students who do not have role models and are questioning whether they have what it takes to excel in an engineering major and career. We focus on grade anomaly as a tool to help understand gender differences in first year engineering grades. Grade anomaly is the difference between a student's grade in a course of interest and their GPA in all other courses thus far excluding that course. We divide grade anomalies into "bonuses" and "penalties". A course in which most students earn a lower grade than usual has a grade penalty, while a course in which most students earn a higher grade than usual has a grade bonus.

We propose that grade anomaly is a potential measure of students' academic self-concept which is easy to track, through institutional grade data. Our framework uses grade penalty as a central construct instead of grade because students' self-concept is tied to what type of student they think they are. Low academic self-concept can be particularly detrimental to women [78–80, 233], so measuring it may be useful for tracking classroom equity. Students tend to have a fairly fixed view of what "kind" of student they are: for example, students may endorse the idea that "If I get As, I must be an A kind of person. If I get a C, I am a C kind of person" [53, 54]. Grade anomalies may challenge or reinforce students' ideas about what kind of student they are, and if they are capable of succeeding in their chosen major. Many students who leave STEM majors explicitly cite lower grades than they are used to as a reason for doing so [53,54]. Grade penalties are more common and extreme in STEM disciplines than in humanities or social science departments [53, 152, 234, 235], and women are more affected by these grade penalties due to stereotypes and lack of role models, and they are more likely to leave their majors or career aspirations with fewer and smaller grade penalties than men are [53, 234].

In this paper, we use Situated Expectancy Value Theory (SEVT), studies about why students leave STEM, and previous work on grade anomalies to explore if grade anomalies in first-year foundational courses affect male and female engineering majors differently, making grade anomalies an equity issue in engineering. We also posit that grade anomaly may be a better measure of self-concept [78] than raw grades and students are more likely to question whether they should continue in disciplines in which the foundational courses involve grade penalty because it is a unique measure of “within-student” frame of reference (i.e., students are comparing their own grades across different courses as opposed to comparing their grades with others) [233].

9.1.1 Prior Work on Grade Penalty

Several studies [152, 235, 238] have utilized “grade anomaly”, the difference between a student’s GPA excluding a course of interest and their grade in all courses thus far. Huberth et al. [238] developed this measure, but Koester et al. [235] conducted the first study we know of that focuses on average grade anomaly (AGA). They used AGA because it was perceived to be a better measure of how students view their comparative performance than their raw grades across different courses. They found that, at their institution, grade penalties were greater for STEM than non-STEM courses. Further, within STEM courses, grade penalties were smaller for men than women. In particular, they found that physics courses had the largest grade penalty and largest gender difference in AGA. The researchers theorized that large grade penalties and gender differences may be partially attributed to high-stakes assessments [13, 94, 120, 121] and stereotype threat [77]. High-stakes assessments like exams are shown to have larger gender differences in grades than low-stakes assessments like problem sets or quizzes [13, 94, 120, 121], while stereotype threat (a students’ feeling of risk associated with confirming a negative stereotype, for example a woman who fears confirming the stereotype that women are bad at math) takes up cognitive resources of students from underrepresented groups [77]. The Matz et al. [152] study had similar findings but with a larger student sample across multiple institutions. Across five universities, STEM courses had larger grade penalties and larger gender differences in AGA that usually favored men. Their study also raised concerns over high-stakes assessments. They emphasized that large grade anomalies often reflect grading decisions made by instructors, rather than being an accurate measure of student learning.

Thus, past work provides evidence for the existence of grade anomalies in STEM courses, and the existence of gender differences in these anomalies. Here, we present an investigation that focuses on grade anomaly in first-year foundational courses in engineering in which we analyze data to study if these trends hold in a more homogeneous population of first-year engineering students at a single university, rather than combining students across institutions and majors. This focus on first-year engineering students can help control for potential confounding factors. This study is particularly important because first year foundational courses in engineering play a critical role in students' short and long-term professional trajectories.

9.1.2 Situated Expectancy Value Theory Framework

Expectancy Value Theory (EVT) [79] and Situated Expectancy Value Theory (SEVT) [78] are frameworks to understand student achievement, persistence, and choice of tasks in a domain (e.g., engineering). EVT posits that performance and persistence is determined by someone's expectation of success and the extent to which they value that task. If a student expects they will succeed in a task and believes that task will be valuable to them (for personal interest, as a path to achieve another goal, etc.) they are more likely to pursue that task. If they do not expect to succeed and do not value a task, they are unlikely to attempt it. Here, we will focus primarily on student expectancies, though value is also important to understanding why some students may persist while others do not.

Expectancies are a combination of academic self-concept, expectations for success, and perceptions of task difficulty [78–80, 233]. Academic self-concept is the most stable and the least task and domain-dependent of the three, and it is based primarily on grades and outside (e.g., from parents, peers, and instructors) feedback [78–80, 233, 236]. Grades inform academic self-concept as both an external ("How good at math am I compared to other students?") and internal ("How good am I at math compared to English?") frame of reference [79, 80, 233].

Expectations of success are more domain and task specific, and refer to a student's belief in their ability to complete a specific task, which will include considerations such as skill in the subject, time allotted, and experience in a subject ([78–80, 233]. Expectancy for success most closely relates to Bandura's theory of self-efficacy [78, 80, 127]. A student may have a positive academic self-concept in math, but may have low expectancy for success if they take a math test on very new material they have not had adequate time to learn. The third expectancy, perceptions of task difficulty, is more straightforward; most students have less faith in their ability to do well on an exam if their peers have reported it to be particularly difficult [79].

In EVT, the three expectancy concepts were collapsed into one factor. However, the updated framework, SEVT, has called to separate these three concepts [78]. According to Eccles and Wigfield [78], combining academic self-concept, expectancies for success, and perceived task difficulty has led to a lack of understanding of the unique developmental mechanisms of each and how the three concepts relate.

We propose that grade anomalies can act as a proxy for student's internal frame of reference. Additionally, past research has found that during times of transition, the usually-stable academic self-concept becomes more dependent on grade feedback and less dependent on outsider (e.g., parental) feedback [233]. We study first-year engineering students because they are more likely to have an unstable academic self-concept due to the transition from high school to university and can be impacted by their performance in first-year courses in college.

9.1.3 Research Questions

We aim to answer the following research questions regarding grade anomalies for first-year engineering students:

RQ1. For which of their first-year courses do engineering students receive a "grade penalty" and for which courses do they receive a "grade bonus"?

RQ2. Do male and female engineering students have different "grade anomalies" in their first-year courses?

RQ3. If there are gender differences in "grade anomalies", do they follow the same trends as gendered grade differences?

9.2 Methods

9.2.1 Participants

This study takes place at the University of Pittsburgh, a large, public, urban, predominantly-white institution in the northeastern United States. The participants were students enrolled in the School of Engineering, who were in their first or second semester at the university, and took calculus-based physics 1 as well as other mandatory first year courses taken by engineering majors between Spring semester of 2006 and Fall semester of 2019. We excluded courses that were taken during the summer semester. This left us with 6,028 engineering majors who took 48,116 courses during their first and second semesters of college. The sample was 29.9% women and 70.1% men. Students who did not list their gender were excluded from the study as they made up less than 0.1% of the sample. Students identified with the following races/ethnicities: 79% White, 9% Asian, 3% Hispanic/Latinx, 3% multiracial, 5% African American/Black, and 1% unknown or unspecified. Demographic data were provided through deidentified university records. This research was carried out in accordance with the principles outlined in the University of Pittsburgh Institutional Review Board (IRB) ethical policy.

9.2.2 Procedures

We chose the courses to include in our investigation by reviewing the engineering first-year curriculum, which is standardized for students at our institution, and confirming that the majority of students took these courses during their first or second semester of college. Chemistry 1 and Chemistry 2 were offered by the Department of Chemistry, but are reserved for engineering students. Physics 1 and Physics 2 were offered by the Department of Physics and Astronomy, and Calculus 1 and Calculus 2 were offered by the Department of Mathematics. All other courses were offered by the school of engineering. Some courses, such as “Composition Seminar”, a writing course for first-semester engineering students, have fewer students than other courses. This is a result of changes in the writing part of the curriculum over the thirteen-year data collection period. All curriculum changes resulted in a very general requirement becoming more specific (for example, students were once required to take one general education course in humanities before graduation, but are now required to take “Composition Seminar” during their first semester). In situations involving these cases, we only included the newer and more specific requirement in our analysis.

9.2.3 Measures

9.2.3.1 Course Grade

Course grades were based on the 0-4 scale used at our university, with A = 4, B = 3, C = 2, D = 1, F = 0 or W (late withdrawal), where the suffixes ‘+’ and ‘-’, respectively, add or subtract 0.25 grade points (e.g., B- = 2.75 and B+ = 3.25), except for the A+, which is reported as 4. We are unable to report detailed grading schemes of each physics instructor, type of course, or any other detailed course-level information but a majority of courses are traditionally taught primarily using lectures.

9.2.3.2 Grade Anomaly

Grade anomaly (GA) was found by first finding each student’s grade point average excluding the course of interest (GPA_{exc}), including all courses taken prior and simultaneously with the course of interest. This was done by using the equation

$$GPA_{exc} = \frac{(GPA_c \times Units_c) - (Grade \times Units)}{Units_c - Units} \tag{5}$$

where GPA_c is the student’s cumulative GPA, $Units_c$ is the cumulative number of units the student has taken, $Grade$ is the grade the student received in an individual course, and $Units$ is the number of units associated with an individual course. After finding GPA_{exc} we can calculate GA by finding the difference between a student's GPA_{exc} and the grade received in that class:

$$GA = Grade - GPA_{exc}. \tag{6}$$

A negative GA corresponds to a course grade lower than a student's GPA in other classes (a "grade penalty"). A positive GA corresponds to a course grade higher than a student's GPA in other classes (a "grade bonus"). Average grade anomaly (AGA) is the mean of students' grade anomalies (GA) for each course, and is the metric by which we compare courses.

9.2.4 Analysis

To characterize both average grade anomaly (AGA) and grades, we found the sample size, mean, standard deviation, and standard error of each measurement for each course of interest. We calculated these statistics for women and men separately, and then for all students combined. We also compared the effect size of gender on both grade and grade anomaly, using Cohen's d to describe the size of the mean differences and unpaired t-tests to evaluate the statistical robustness of the differences. Cohen's d is calculated as follows:

$$d = \frac{\mu_1 - \mu_2}{\sqrt{\sigma_1^2 - \sigma_2^2}} \tag{7}$$

where μ_1 and μ_1 are the means of the two groups, σ_1 and σ_2 are the standard deviations [108] and Cohen's d is considered small if $d \sim 0.2$, medium if $d \sim 0.5$, and large if $d \sim 0.8$ [140]. We used a significance level of 0.05 in the t-tests and as a balance between Type I (falsely rejecting a null hypothesis) and Type II (falsely accepting a null hypothesis) errors [108]. All analysis was conducted using R [174], using the package plotrix [245] for descriptive statistics, lsr [246] for effect sizes, and ggplot2 [247] to create plots.

9.3 Results

9.3.1 For which of their first-year courses do engineering students receive a "grade penalty" and for which do they receive a "grade bonus"?

To answer RQ1, we calculated average grade anomaly (AGA) for each course engineering students are required to take during their first and second semester at the university. We show the descriptive statistics for both grades and AGA in Table 35 and Figure 12. The largest student sample can be found for Physics 1, because this is the class we used to select engineering students for our sample. We find that students generally receive grade penalties in the courses offered by the departments of Chemistry, Mathematics, and Physics, while students receive grade bonuses in first-year courses offered by the department of English and School of Engineering. The courses in which students receive a grade penalty are (in order from largest to smallest penalty): Chemistry 1, Chemistry 2 and Physics 2 (tie), Calculus 2 and Physics 1 (tie), and Calculus 1. The courses that students receive a grade bonus are (in order from smallest to largest bonus): Introduction to Computing, Composition Seminar, Introduction to Analysis, and Engineering Communication in a Professional Context.

9.3.2 Do male and female engineering students have different "grade anomalies" in their first-year courses?

To find if there are differences in grade anomalies for men and women, we grouped students by their self-reported gender and calculated the average grade anomaly for both groups for each course of interest. We then calculated Cohen's d as a measure of effect size between the two groups [108], which can be seen in Table 36. Women had indistinguishable or favorable AGA outcomes (e.g., smaller grade penalties or larger grade bonuses) compared to men, with three exceptions (see Figure 13). For Intro to Computing, Physics 1, and Physics 2, men have smaller grade penalties or larger grade bonuses than women.

For both men and women, Professional Communication, a writing course, provided the largest grade bonus. For men, the courses that provided the largest grade penalties are Chemistry 1 and Chemistry 2, with grade penalties of -0.62 and -0.64, respectively. This means that men tended to receive over half a letter grade lower in this course than in their other courses. For women, the course that provided the largest grade penalty is Physics 2, with an AGA of -0.71.

9.3.3 4.3. If there are gender differences in "grade anomalies", do they follow the same trends as gendered grade differences?

There are some classes that show similar trends for grades and AGA, which can be seen in Table 36. We define similar trends as having a similar effect size (small, medium, or large) and a similar p-value. These include Composition Seminar, Chemistry 1, Chemistry 2, Calculus 1, and Calculus 2. There are courses with a larger gender difference in grade than AGA. These include Intro to Computing, Intro to Analysis, and Engineering Communication in a Professional Context. There are courses with a larger gender difference in AGA than grade. These include Physics 1 and Physics 2.

9.4 Discussion

Our results show that there are grade penalties in all science and math courses studied, and bonuses in all engineering and English courses. We note that, similar to other studies that focus on AGA [152, 235], science and math courses have large grade penalties, while humanities courses have grade bonuses. Our results differ from past studies because first year courses offered by the engineering school have grade bonuses as opposed to penalties. In this section, we discuss: the potential harms of grade anomalies, what gender differences in grade anomalies can reveal about course equity, how grade anomaly related to academic self-concept; concerns about unequal access to coping mechanisms (methods that students use to persist in an environment with large grade penalties) regarding grade penalties, and why grade anomalies are a useful measure above and beyond raw grades.

First, we discuss why grade anomalies can be harmful. Lower than expected grades, even in a single course, can be a catalyst for students to leave STEM majors [53, 54]. This does not just include D and F grades or withdrawal from the course, but grades that were high enough to continue the program that did not meet a student's personal expectations [53, 54]. This was a particular issue among high-achieving students, who were more likely to endorse perfectionism and feeling that their identity as "good STEM students" was threatened by B's and C's, or even a low grade on a single exam [53].

One way that students report coping with these unexpectedly low grades is by relying on others (such as friends, professors, or mentors) for support, which often comes as reassurance that low grades are normal in these difficult classes [53]. A second way students report coping is by accepting the harsher grading standards (such as curved grading or very low class exam averages) of STEM courses and decoupling their self-concept as STEM students with their grades [53]. Both of these coping mechanisms raise equity concerns, which are discussed in the next section.

It is important to note that these grading standards are a choice made by STEM departments and instructors. Requiring students to accept harsh grading standards and separate their identities as STEM students from their grades in order to successfully complete their degrees can distract students from their coursework. This, combined with evidence that shows that many high-achieving students leave these majors due to grade-related concerns, should lead instructors to question if their standards actually improve the education they offer students, or if they are simply pushing away all students except those who are capable of maintaining their academic self-concept divorced from grades.

In addition to seeing evidence of grade penalties in some courses for first-year engineering students, especially large-enrollment introductory courses, we also see some gender differences in grade anomalies. In particular, there were larger grade penalties for women in Physics 1 and Physics 2. Because women leave majors with higher grades than men who remain both in [44, 53] and outside [234] of STEM, this raises serious equity concerns. Past work suggests that women tend to have lower motivational beliefs that relate to academic self-concept, such as self-efficacy and sense of belonging [10, 12, 41, 44, 94, 120, 224, 241, 242]. Women report feeling more demoralized than men when they receive low grades, and cite more worry over not understanding material even if they receive A's, B's, or C's (all of which are grades that allow students to continue in most programs) [53, 115]. This trend has been found to be particularly strong among high-achieving women [53].

We suggest that women may be more likely to have a low academic-self-concept than men at similar performance levels for two reasons. First, prior research also suggests that women are less likely to separate their academic self-concept from their grades which is one of the clearest types of recognition in a domain [53, 54, 234]. In particular, grades are the resource that women have the most access to. Academic self-concept is formed through grades and feedback from outsiders. Because women are less likely to receive recognition as someone with potential in STEM from their parents [67, 97, 181], society at large [55, 57], and their instructors [56, 130, 208], they are more likely to rely on grade information to develop their academic self-concept. Next, women often tend to earn higher grades than men with the same standardized test scores [53,203]. Because women are often more accustomed to higher grades, they may have more concern about grades that are lower than what they are accustomed to (especially if there are stereotypes about who can excel in those domains), or they may compare their relatively-low STEM grades and leave for another subject that gives them the recognition for their work that they are accustomed to [53, 54].

Next, interview-based studies [53] suggest that women are less likely to have access to the coping mechanisms (i.e., support from peers or mentors and resources to decouple their self-identity as STEM students with their grades [53]) that students often use to continue even if they receive lower-than-expected grades in foundational courses. For example, women are less likely to receive advice that low grades are acceptable from peers and mentors because they are less likely to have peers and mentors they can relate to due to the underrepresentation of women in many STEM fields, such as engineering [1]. Further, if women are less likely to continue in their field of interest due to low grades [53, 115], the women who remain in or complete programs are more likely to be high-achieving in the field, and would thus be less likely to give advice that is useful to the average student (for example "I, like many others, received a C in this class but was still able to complete my program"). This second coping mechanism is essentially separating grades from academic self-concept. As stated earlier, because women are less likely to have positive feedback from outsiders (e.g., parents and instructors), they have no other way to form academic self-concept unless they can find a support system that can provide that outside recognition, though this sort of system may not be available to every student who seeks it out. These same arguments about lack of access to coping mechanism may apply to students from other underrepresented groups, who are also more likely to leave their programs due to lower-than-expected grades, i.e., due to grade penalty, than students from groups that are not underrepresented. Other groups that may be more affected by grade anomalies are racial and ethnic minority students [47, 53] and first-generation students [248].

Finally, we find that grade anomalies and raw grade data do not always reveal the same trends. Some courses have larger gender differences in AGA than in grades, such as Physics 1 and Physics 2. This speaks to the usefulness of tracking both AGA and grades of the students. An instructor may see a small grade difference and understand that there is gender inequity in their classroom, but without knowing the gender differences in AGA, the instructor will not understand how those grades are perceived by female and male students. Understanding both grades and AGA differences may allow instructors to understand both classroom-level inequities and the extent to which their course may be pushing underrepresented groups, such as women, out of STEM fields.

There are also some courses that have larger gender differences in grades than in AGA, such as Intro to Computing, Intro to Analysis, and Engineering Communication in a Professional Context. However, we do not find these differences as concerning, because all of these courses are, on average, offering grade bonuses to all students, so they are less likely to decrease students' academic self-concept.

From our results, we make several recommendations to instructors and departments. First, measuring grade anomaly in addition to grades may be a useful way to find inequities in the learning environment. Measuring grades and gendered grade differences is both valuable for and accessible to individual instructors, but grade anomalies may be useful to departments concerned about students' retention over longer periods and finding which courses may be discouraging students from under-represented groups to leave a major.

Next, we encourage instructors and departments to evaluate their goals when developing grading practices. If a department aims to create a diverse and welcoming environment that attracts students, while also maximizing student learning, there are several productive practices to consider. Frequent, low-stakes assessment gives ample opportunity for instructor feedback and can minimize gender inequities in STEM classrooms [94, 120, 127]. This includes offering many types of assessment, such as quizzes, clicker questions, and projects in addition to or instead of homework problem sets and exams [13, 94, 119]. Collaborative and active learning approaches in equitable learning environments may also improve learning outcomes and grades while eliminating gender performance differences [94,120,157,249]. Finally, we recommend instructors avoid curved grading and very low class averages: these practices do not reflect student performance, but they do discourage students and often contribute to students' reasons for leaving a field [53, 54].

The results presented in this study are very important because they provide evidence that courses in STEM departments (particularly large, mandatory, introductory courses) tend to result in grade penalties for students. This allows us to pinpoint departments and courses that may have grading practices that are inequitable or unrepresentative of student learning, as well as those that have more equitable and representative grading, so that practices may be shared across disciplines to improve learning environments in all disciplines. The relatively new measure of AGA may also act as a measure of academic self-concept that is easy for institutions to access. This can also be useful to researchers as they develop separate measurements for academic self-concept and expectancies for success. Because most research on measures of self-concept is relatively new [78], qualitative work in this area may help further clarify the connection between grade anomalies and academic self-concept, as well as reveal how they both affect retention. Further, grade anomaly may correlate with multiple factors, not just self-concept (for example, student self-efficacy, interest, course engagement, and impact of teaching methods), and qualitative or survey data may reveal more nuanced impact for each of these factors on grade penalty.

Although we have strong evidence of grade penalties in chemistry, mathematics and physics courses for first-year engineering students as well as gendered grade anomaly differences, we did not have access to syllabi or other information about individual courses for every course offered over the thirteen-year period of data collection. Therefore, we are not able to pinpoint specific practices that may lead to grade penalties, grade bonuses, or gender inequities at our institution. Instead, we assume that, like the courses currently offered, most of these large, introductory courses are taught in a traditional, lecture-based, and exam-reliant format.

9.5 Conclusion

In this work we found that grade anomalies exist for all first-year engineering courses at our institution. Engineering and English Composition courses offered grade bonuses while Physics, Math, and Chemistry courses had grade penalties. Further, all courses had a grade anomaly (larger grade bonuses or smaller grade paneities) that favored women over men except for both Physics courses and Introduction to Computing. This raises particular concern about physics classes and the need for an equitable learning environment for engineering students. We also note that grade anomalies and raw grades do not reveal the same gender difference trends. Thus, both grade anomaly and raw grades should be tracked when determining if a course is equitable.

Finally, this research is based at a primarily white, large, public university, and while our results may generalize to similar institutions, we do not know what patterns of grade anomalies exist at liberal arts colleges, minority-serving institutions, or community colleges. Conducting research at a diverse range of institutions in different countries, as well as a stronger focus on how grade anomaly affects students from a variety of underrepresented groups, will help us more fully understand how grade anomalies differ for a range of students.

Table 35: Grades and AGA for each course of interest, including the department that offered it and semester in which it was offered. The following words are abbreviated: Seminar (Sem.), Computing (Comp.), Communication (Com.), Professional (Prof.), and Standard Deviation (SD).

Course	Semester	Department	N	Mean	SD	Mean	SD
Composition Seminar	Fall	English	787	3.51	0.72	0.48	0.70
Intro to Analysis	Fall	Engineering	5352	3.52	0.53	0.55	0.54
Chemistry 1	Fall	Engineering	4185	2.55	0.99	-0.62	0.85
Physics 1	Fall	Physics	6022	2.73	0.79	-0.47	0.72
Calculus 1	Fall	Mathematics	4381	2.81	1.01	-0.23	1.08
Intro to Comp.	Spring	Engineering	4672	3.29	0.73	0.27	0.62
Prof. Com.	Spring	English	338	3.81	0.31	0.63	0.55
Chemistry 2	Spring	Chemistry	3385	2.52	0.91	-0.59	0.68
Physics 2	Spring	Physics	4762	2.58	0.91	-0.58	0.70
Calculus 2	Spring	Mathematics	4601	2.63	1.13	-0.48	0.99

Table 36: Comparison of grades and AGA between men and women. If the effect size given by Cohen's d is positive for grade, women had higher grades than men. If d is positive for AGA, women had a larger grade bonus or smaller grade penalty than men in that course. The following words are abbreviated: Seminar (Sem.), Computing (Comp.), Communication (Com.), Professional (Prof.), and Standard deviation (SD). $c = p < 0.05$, $b = p < 0.01$, and $a = p < 0.001$

Course	N	Mean	SD	Mean	SD	N	Mean	SD	Mean	SD	Grade	AGA
Composition Sem.	277	3.70	0.48	0.64	0.65	510	3.41	0.80	0.40	0.71	0.41^a	0.34^a
Intro to Comp.	1352	3.34	0.70	0.24	0.57	3320	3.26	0.75	0.28	0.63	0.10^a	-0.05
Intro to Analysis	1580	3.60	0.49	0.55	0.47	3772	3.49	0.55	0.55	0.57	0.20^a	0.00
Prof. Com.	129	3.86	0.35	0.68	0.52	209	3.77	0.34	0.60	0.57	0.26^a	-0.10
Chemistry 1	1103	2.68	0.90	-0.55	0.81	3082	2.50	0.98	-0.64	0.86	0.19^b	0.10^b
Chemistry 2	899	2.67	0.84	-0.50	0.63	2486	2.46	0.93	-0.62	0.69	0.22^a	0.17^a
Physics 1	1783	2.67	0.73	-0.65	0.68	4239	2.76	0.81	-0.40	0.72	-0.11^a	-0.36^a
Physics 2	1342	2.54	0.82	-0.71	0.64	3420	2.59	0.94	-0.54	0.71	-0.06^c	-0.25^a
Calculus 1	1222	2.98	0.90	-0.10	0.99	3159	2.75	1.04	-0.28	1.11	0.23^a	0.16^a
Calculus 2	1288	2.79	1.06	-0.38	0.93	3313	2.63	1.13	-0.48	0.99	0.20^a	0.15^a

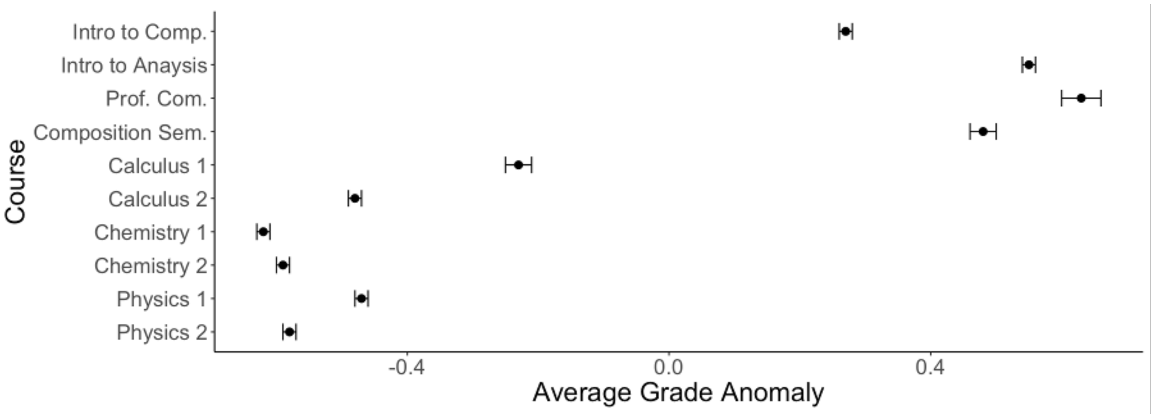


Figure 12: Average grade anomaly (AGA) of all students by course. The following words are abbreviated: Seminar (Sem.), Computing (Comp.), Communication (Com.), and Professional (Prof.). Ranges represent standard error of the mean.

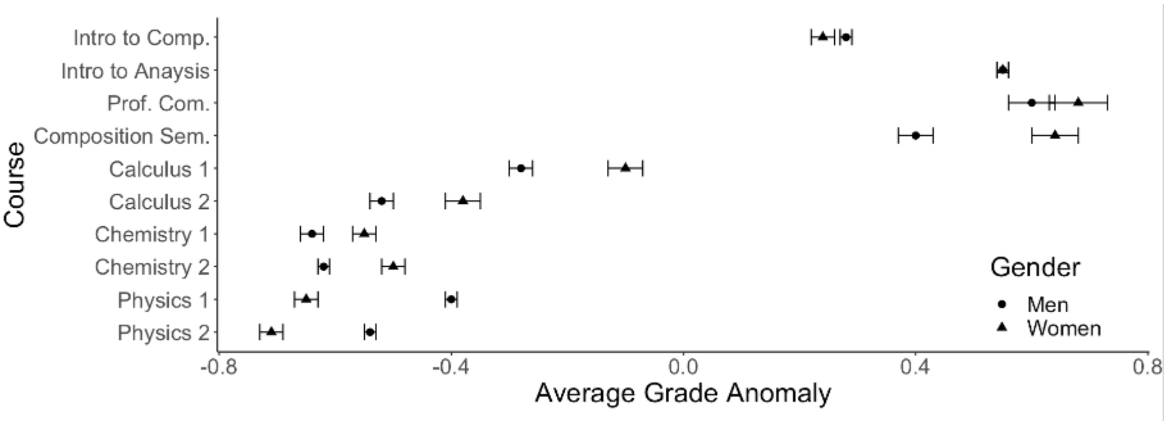


Figure 13: Comparison of average grade anomaly (AGA) between men and women. The following words are abbreviated: Seminar (Sem.), Computing (Comp.), Communication (Com.), Professional (Prof.), and Standard deviation (SD). Ranges represent standard error of the mean.

10.0 Gender gaps in grades versus grade penalties: Why grade anomalies may be more detrimental for women aspiring for careers in biological sciences and contribute to a leaky pipeline

10.1 Introduction

Research on gendered performance and persistence differences in science, technology, engineering, and mathematics (STEM) fields is important in the fields in which women are the least likely to earn degrees, such as physics or engineering [4, 7, 11, 12, 41, 43, 45, 65, 115]. However, less research has been conducted on gender differences in fields in which women are not underrepresented, such as biology [120, 240]. Concerns about gender equity in STEM education are not limited to the number of women in the field, but includes the experiences of female students when they do participate [24, 25, 56, 188, 250]. In domains such as biological science, woman may earn the same number of undergraduate degrees as men [1], even if men and women have different experiences in the classroom.

Prior research has found that women are less likely to participate in class discussions in biology classrooms [251] and are less likely to be viewed as knowledgeable by their peers [182]. Further, women and tend to have lower exam grades than men in their introductory biology classes [94]. Despite their different classroom experiences, women earn undergraduate and graduate bioscience degrees at higher rates then men [1]. Becuase biology degree recipients make up over half of medical school applicants and matriculants in the United States [252], it is also important to note that there are more women from 2018-2021 who entered medical school than men [252]. If women are earning degrees at high rates regardless, why would it be important to investigate undergraduate classroom inequities?

The answer may lay in gender-differences after graduation. Women who earn biological science degrees are less likely than men to work as scientists after receiving graduate degrees [253]. If they do continue in their field of study, women who pursue medical careers or academic biology are likely to experience gender-based inequities. For example, there are gender disparities in compensation and time to promotion for all academic medical specialties [254] as well as for physicians [255]. In addition, one recent study [256] found that in biology, women tend to have shorter publishing careers (due to leaving the field) than men, and during those careers men had higher annual publishing rates than women. The annual publishing rate for Biology had a larger gender difference favoring men than any other domain studied, including those in which women are underrepresented at all levels, such as physics and computer science [256]. Other work has estimated that Biology authorship will not reach gender parity for another twenty-five years [217]. Despite their lower publishing rates, women in biology also feeling more stress arising from the pressure to publish than men do [257].

If gender differences career outcomes are explained a lack of representation in the classroom, than motivational factors may provide some insight. One such motivational construct is academic self-concept, which describes a long-term expectation of success that students hold regarding their academic abilities and that depends on outside feedback, such as grades [79,80,233]. Low academic self-concept may lead to lower future achievement and persistence because it discourages student engagement in a domain [79]. When women leave STEM disciplines, they often do so with higher grades than the men who remain in the program [39,53,54].

There are many potential partial explanations that have been suggested regarding why women who are meeting or exceeding the requirements of their programs leave, or do not continue after graduation. These include lack of societal biases about who can excel in these disciplines [56, 57, 59, 67, 202], gender discrimination in hiring [208], and differences in STEM motivational beliefs such as self-efficacy and sense of belonging [6–8, 11, 12, 41, 65, 115, 188]. One related reason for why this happens may be lower academic self-concept of female students in these courses compared to male students. Though none of these factors may provide complete explanations of gender differences, aiming to address them simultaneously may create a better learning environment for women.

Here, we focus on bioscience majors and inquire about gender differences in grade penalties. In order to quantify grade penalty, we define grade anomaly as the difference between a student's grade in a course of interest and their grade point average (GPA) in all other classes up to that point. The mean of this statistic for all students who took a course is the average grade anomaly (AGA). We divide average grade anomalies into "bonuses" and "penalties". A course in which students on average earn a lower grade than usual has an AGA with grade penalty, while a course in which students on average earn a higher grade than usual has an AGA with grade bonus.

Within our framework, we posit that grade anomaly may allow us to track, through institutional grade data, an important measure of how courses may affect students' academic self-concept. Our framework uses grade penalty as a central construct instead of grade because students' academic self-concept is often based on comparisons, not absolute grades [78]. Students may compare their grades across courses to determine which disciplines they excel at or struggle with [78]. Additionally, students tend to have a fairly fixed view of what "kind" of student they are, e.g., students may endorse the idea that "If I get As, I must be an A kind of person. If I get a C, I am a C kind of person" [53]. Grade anomalies may challenge or reinforce students' ideas about what kind of student they are, and if they are capable of succeeding in their chosen domain. Many students who leave STEM explicitly cite lower grades than they are used to as a reason for doing so [53, 54]. Grade penalties are more common and extreme in STEM disciplines than in humanities or social science departments [53, 152, 234, 235], and women tend to have larger grade penalties than men in STEM subjects [152].

In this paper, we use Situated Expectancy Value Theory (SEVT), studies about why students leave STEM, and previous work on grade anomalies to explore whether the average grade anomalies for men and women in biology and related majors are different, making grade anomalies an equity issue. We also posit that grade anomaly may be a better measure of self-concept than raw grades because it is a measure of a "within-student" frame of reference (i.e., students are comparing their own grades across different courses as opposed to comparing their grades with others) [78, 233].

10.1.1 Research Questions

We aim to answer the following research questions regarding grade anomalies:

RQ1. For which of their courses do students majoring in biological sciences receive a "grade penalty" and for which courses do they receive a "grade bonus"?

RQ2. Do men and women have different "grade anomalies" in their STEM courses?

RQ3. Do gender differences in "grade anomalies" follow the same trends as gender differences in grades?

10.1.2 Theoretical Framework

10.1.2.1 Expectancy Value Theory

Expectancy Value Theory (EVT) [79] and Situated Expectancy Value Theory (SEVT) [78] are frameworks to understand student achievement, persistence, and choice of tasks in a domain (e.g., biology or neuroscience). EVT posits that performance and persistence are determined by someone's expectation of success and the extent to which they value that task. If a student expects they will succeed in a task and believes that the task will be valuable to them (for personal interest, as a path to achieve another goal, etc.) they are more likely to pursue that task. If they do not expect to succeed and do not value a task, they are unlikely to attempt it. Here, we will focus primarily on student expectancies, though value is also important to understanding why some students may persist while others do not. Expectancies are a combination of academic self-concept, expectations for success, and perceptions of task difficulty [78–80, 233]. Academic self-concept is the most stable and the least task and domain-dependent of the three, and it is based primarily on grades and outside (e.g., from parents, peers, and instructors) feedback [78–80, 233, 236]. Grades inform academic self-concept as both an external ("How good at math am I compared to other students?") and internal ("How good am I at math compared to English?") frame of reference [79, 80, 233].

Expectancies of success are more domain and task specific and refer to a student's belief in their ability to complete a specific task, which will include considerations such as knowledge and skill related to the subject, time allotted, and experience in a subject [78–80, 233]. Expectancy for success most closely relates to Bandura's theory of self-efficacy [78, 80, 127]. A student may have a positive academic self-concept in math, but may have low expectancy for success if they take a math test on very new material they have not had adequate time to learn. The third expectancy, perceptions of task difficulty, is more straightforward; most students have less faith in their ability to do well on an exam if their peers have reported it to be particularly difficult [79].

In EVT, the three expectancy concepts were collapsed into one factor. However, the updated framework, SEVT, has called for a separation of these three concepts [78]. According to Eccles and Wigfield, combining academic self-concept, expectancies for success, and perceived task difficulty has led to a lack of understanding of the unique developmental mechanisms of each and how the three concepts relate [78]. We posit that grade anomaly may be a better measure of self-concept [78] than raw grades. This is because students often judge their ability by comparing their grades across courses rather than comparing their grades to other students' (in EVT/SVET, this is called the "within student" frame of reference). Poor performance from a within-student frame of reference may cause students to question if they should continue in a discipline [78].

10.1.2.2 Grade Anomalies

Grade anomalies allow us to measure how courses can affect students' academic self-concept [78, 80, 234] using institutional grade data [152], which may be more accessible to instructors and researchers than surveys or interview data. While students' raw grades are a useful measure because they allow for direct comparison between students and because they are used by institutions to award scholarships and track student academic standing, we propose that using grade penalty in addition to raw grades gives researchers and instructors more insight into student self-concept (that is, a students' view of what "kind" of student they are). This is because students tend to have a fairly fixed self-concept and often endorse the idea that "If I get As, I must be an A kind of person. If I get a C, I am a C kind of person" [53].

A student who receives lower grades in their science courses than their humanities courses may take this as a sign that they are not capable of excelling in the sciences, even if the grades they earn are high enough for them to continue in their major [53, 54]. This experience may be common, because grade penalties tend to be more extreme and widespread in STEM disciplines than in other subjects [53,235,237,238].

Several studies [152, 235, 239] have utilized "grade anomaly" or "grade penalty", the difference between a student's GPA excluding a course of interest and their grade in all courses thus far. Koester et al. [235] conducted the first study we know of that focuses on average grade anomaly (AGA). They used AGA because it was perceived to be a better measure of how students view their comparative performance than their raw grades across different courses. They found that, at their institution, grade penalties were greater for STEM than non-STEM courses. Further, within STEM courses, grade penalties were smaller for men than women. In particular, they found that physics courses had the largest grade penalty and largest gender difference in AGA. The researchers theorized that large grade penalties and gender differences may be partially attributed to high-stakes assessments [13, 94, 120, 121, 240], and stereotype threat [77]. The Matz et al. [152] study had similar findings but with a larger student sample across multiple institutions. Across five universities, STEM courses had larger grade penalties and larger gender differences in AGA that usually favored men.

Thus, past work provides evidence for the existence of grade anomalies in STEM courses, and the existence of gender differences in these anomalies. Here, we present an investigation that focuses on grade anomaly in various courses for biological science majors in which we analyze data to study if these trends hold in a more homogeneous population of largely pre-health and pre-medical students at a single large university in the US, rather than combining students across institutions and many majors.

10.2 Methodology

10.2.1 Participants

Participants in this study were enrolled in bioscience or health majors at the University of Pittsburgh, which is a large, public, and urban institution. The student major breakdown was as follows: 37% Biological Sciences, 1% Bioinformatics/Computational Biology, 3% Ecology and Evolution, 6% Microbiology, 6% Molecular Biology, 30% Neuroscience, 4% Pharmacy, and 13% Rehabilitation Science. All major except for Neuroscience, Pharmacy, and Rehabilitation Science are offered through the Department of Biological Sciences. These students were chosen because of their similar course requirements, especially for large introductory science courses.

Grade data collected over thirteen years, and we excluded courses that were taken during the summer semester. We excluded summer courses because they are not a typical representation of courses at our institution. For example, many summer students do not primarily attend our institution, but are local students visiting home for the summer. In addition the class sizes are an order of magnitude smaller than those in the Fall and Spring semesters. This left us with 2,445 students who took 89,560 courses. The sample was 58.1% women and 41.9% men. Less than 0.1% did not list their gender, so they were excluded from the study due to small sample size.

Students identified with the following races/ethnicities: 68% White, 17% Asian, 3% Hispanic/Latinx, 3% multiracial, 7% African American/Black, and 1% unknown or unspecified. This research was carried out in accordance with the principles outlined in the University of Pittsburgh Institutional Review Board (IRB) ethical policy, and de-identified demographic data were provided through university records.

10.2.2 Course Selection

We chose to study courses that were taken by the largest number of students, excluding non-major electives (for example, “Introduction to Piano” or ”Public Speaking”) and courses that make up general education requirements. Thus, many courses were mandatory for students in the majors we focus on. However, not all courses were required for students in all the majors in our sample. Information about if a course was required, optional (i.e., an elective that count towards the major), or not required is included in Table 37. The courses we chose are listed in Table 38, along with information about the year in which the students typically take the course. We would like to note that, though it is not required for most majors studied, Human Physiology met our criteria because it is a commonly-chosen elective for both Biology and Rehabilitation Science Students. In addition, students in this sample may take either calculus or algebra-based physics, but so few ($N = 61$) students chose calculus-based physics that they were excluded from this study.

10.2.3 Measures

10.2.3.1 Course Grade

Course grades were based on the 0-4 scale used at our university, with A = 4, B = 3, C = 2, D = 1, F = 0 or W (late withdrawal); the suffixes '+' and '-', respectively, add or subtract 0.25 grade points (e.g., B- = 2.75 and B+ = 3.25), except for the A+, which is reported as 4. We are unable to report grading schemes of each instructor, type of course (i.e., traditional lectures or active learning), or any other detailed course-level information due to the large number of courses sampled.

10.2.3.2 Grade Anomaly

GA was found by first finding each student's grade point average excluding the course of interest (GPA_{exc}). This was done by using the equation

$$GPA_{exc} = \frac{GPA_c \times Units_c - Grade \times Units}{Units_c - Units} \tag{8}$$

where GPA_c is the student's cumulative GPA, $Units_c$ is the cumulative number of units the student has taken, $Grade$ is the grade the student received in an individual course, and $Units$ is the number of units associated with an individual course. After finding GPA_{exc} we can calculate grade anomaly (GA) by finding the difference between a student's GPA_{exc} and the grade received in that class:

$$GA = Grade - GPA_{exc}. \tag{9}$$

A negative GA corresponds to a course grade lower than a students' GPA in other classes and we call this a "grade penalty". A positive GA corresponds to a course grade higher than a students' GPA in other classes and we call this a "grade bonus". Average grade anomaly (AGA) is the mean of students' grade anomalies (GA) for each course, and is the metric by which we compare courses.

10.2.3.3 Analysis

To characterize both average grade anomaly (AGA) and grades, we found the sample size, mean, standard deviation, and standard error of each measurement for each course of interest. We calculated these statistics for women and men separately, and then for all students combined. We also compared the effect size of gender on both grade and grade anomaly, using Cohen's d to describe the size of the mean differences and unpaired t-tests to evaluate the statistical robustness of the differences. Cohen's d is calculated as follows:

$$d = \frac{\mu_1 - \mu_2}{\sqrt{\sigma_1^2 - \sigma_2^2}} \tag{10}$$

where μ_1 and μ_1 are the means of the two groups, σ_1 and σ_2 are the standard deviations [108] and Cohen's d is considered small if $d \sim 0.2$, medium if $d \sim 0.5$, and large if $d \sim 0.8$ [140]. We used a significance level of 0.05 in the t-tests and as a balance between Type I (falsely rejecting a null hypothesis) and Type II (falsely accepting a null hypothesis) errors [108]. All analysis was conducted using R [174], using the package plotrix [245] for descriptive statistics, lsr [246] for effect sizes, and ggplot2 [247] to create plots.

10.3 Results

10.3.1 For which of their courses do bioscience students receive a “grade penalty” and for which courses do they receive a “grade bonus”?

To answer RQ1, we calculated average grade anomaly (AGA) for the most popular courses taken by biological science students and students in related majors focusing on future careers in health professions. We show the descriptive statistics for both grades and AGA in Table 39 and Figure 14. We find that students generally received grade penalties in all STEM courses we studied. One course, Human Physiology, has a smaller grade penalty than other courses, while Organic Chemistry 1 and 2 have much larger grade penalties than all other courses. For Organic Chemistry courses, students have an AGA of approximately 1, meaning on average, students receive one full letter grade lower in these courses than in their other courses. There are no grade bonuses in the table because not all courses students take are included. For example, students may receive grade bonuses in general education courses, e.g., in social sciences and humanities as well as in laboratory courses.

10.3.2 Do men and women have different “grade anomalies” in their STEM courses?

To investigate if there are differences in grade anomalies between men and women, we grouped students by their self-reported gender and calculated the average grade anomaly for both groups for each course of interest. We then calculated Cohen’s d as a measure of effect size between the two groups [108], which can be seen in Table 40. Group differences can also be seen in Figure 15. Women had statistically indistinguishable AGA outcomes to men in Calculus, Human Physiology, Genetics, Biochemistry, and Chemistry 2. By indistinguishable, we mean there is no statistically significant difference between men and women's AGAs in Table 40. Women did not have favorable AGA outcomes (e.g., smaller grade penalties) compared to men in any course. Men had favorable AGA outcomes compared to women in several courses, meaning there was a statistically significant difference in AGA between man and women, and men tended to have smaller grade penalties than women. These courses included Biology 1, Biology 2, Organic Chemistry 1, Organic Chemistry 2, Chemistry 1, Physics 1, and Physics 2.

10.3.3 Do gender differences in "grade anomalies" follow the same trends as gender differences in grades?

There are many courses for which there was no statistically significant differences in grades or grade anomalies, such as Human Physiology, Genetics, Biochemistry, Chemistry 2, and Physics 2, which can be seen in Table 40. With the exception of Chemistry 2, these are all courses that students tend to take in their second year of university or later. There was also one course in which the gender differences were similar for grades and AGA: Organics Chemistry 1. Cohen's d between genderes was similar for Grade and AGA, as shown in Table 40. These similar but significant gaps favoring men can also be seen in Figures 15 and 16. For the aforementioned six courses, AGA and grades provide similar information.

There are also courses that show difference trends in AGA versus raw grades. For example, Table 40 reveals that Biology 1, Biology 2, Organic Chemistry 2, and Chemistry 1 have no statistically significant gender difference in grades, but all have a statistically significant difference between men and women when in AGA. Similarly, Table 40 also shows that the gender difference in AGA ($d = 0.26$) is larger than in raw grades ($d = 0.11$) for Physics 1, even if the gender difference is statistically significant for both measures. Comparing Figures 15 and 16, it is clear that the gender differences are larger for AGA than raw grades for all of these courses.

There is one course that has a larger gender difference in grades than in AGA: Calculus 1, which can be seen in Table 40. In this course, women tend to have indistinguishable AGAs to men, but statistically significantly higher grades than men.

10.4 Discussion

Our results show that there are grade penalties in all courses studied. First, we discuss why grade anomalies can be harmful. Lower than expected grades, even in a single course, can be a catalyst for students to leave STEM majors [53, 54]. This does not just include D and F grades or withdrawal from the course, but grades that were high enough to continue the program that did not meet a student's personal expectations [53,54]. This was a particular issue among high-achieving students, who were more likely to endorse perfectionism and feeling like their identity as "good STEM students" was threatened by B's and C's, or even a low grade on a single exam [53]. Thus, we believe that the courses that have the largest grade penalties, in this case Organic Chemistry 1 and 2, are the courses most likely to push students out of these disciplines or cause them to question their abilities.

In addition to seeing evidence of grade penalties in some courses we also see evidence of gender differences in grade anomalies in over half of the courses studied, particularly Biology 1 and 2, Organic Chemistry 1 and 2, Chemistry 1, and Physics 1 and 2. In particular, we find Organic Chemistry 1 and 2 concerning, because this is the largest grade penalty women receive, so it is likely to stand out as a unique anomaly, and may affect women's academic self-concept more than other courses. Women report feeling more demoralized than men when they receive low grades, and cite more worry over not understanding material even if they receive A's, B's, or C's (all of which are grades that allow students to continue in most programs) [53, 115]. This trend has been found to be particularly strong among high-achieving women [53].

We hypothesize that women may be more likely to have a low academic-self-concept than men at similar performance levels for two reasons. First, prior research suggests that women are less likely to separate their academic self-concept from their grades, which is one of the clearest types of recognition in a domain [53, 54, 237]. In particular, grades are the resource that women have the most access to. Academic self-concept is formed through grades and feedback from outsiders. Because women are less likely to receive recognition as someone with potential in STEM from their parents [67, 97, 181], society at large [55, 57], and their instructors [56, 130, 208], they are more likely to rely on grade information to develop their academic self-concept. Next, women often tend to earn higher grades than men with the same standardized test scores [53,203]. Because women are often more accustomed to higher grades, they may have more concern about grades that are lower than what they are accustomed to, or they may compare their relatively-low STEM grades and view themselves as less able to succeed in biology than a subject that gives them the recognition for their work that they are accustomed to [53, 54].

Finally, we find that grade anomalies and raw grade data do not always reveal the same trends. Some courses have larger gender differences in AGA than in grades, such as Biology 1 and 2, Chemistry 1, and Physics 1. This speaks to the usefulness of tracking both AGA and grades of the students. An instructor may see a small or negligible grade difference by gender and assume that there is gender equity in their classroom based upon grade outcomes by gender, but without knowing the gender differences in AGA, the instructor will not understand how those grades are perceived by female and male students. Understanding both grades and AGA differences may allow instructors to understand both classroom-level inequities.

There is one course that has larger gender differences in grades than in AGA: Calculus 1. We find this course concerning because of the average grade penalty for all students, particularly because this course is most often taken by first-year students, who are more likely to have an academic self-concept that is in flux [78].

Measuring grade anomaly in addition to grades may be a useful way to find inequities in the learning environment. Measuring grades and gendered grade differences is both valuable for and accessible to individual instructors, but grade anomalies may be useful to departments concerned about students' retention over longer periods and finding which courses may be particularly discouraging to students from underrepresented groups.

10.5 Conclusion and Future Research

In this work we found that grade penalties exist for all the courses we studied. Further, seven courses had a grade anomaly (larger grade bonuses or smaller grade penalties) that favored men over women, while five courses had a grade anomaly that did not favor either gender. This raises particular concern about the need for an equitable learning environment for these students.

These results are very important because they provide evidence that courses in STEM departments tend to have grade penalties. This support the results found in prior work on grade anomalies in a more homogeneous population. The relatively new measure of AGA may also act as a measure of academic self-concept that is easy for institutions to access. This can also be useful to researchers as they develop separate measurements for academic self-concept and expectancies for success.

Although we have evidence of grade penalties in the studied courses as well as gendered grade anomaly differences, we did not have access to syllabi or other information about individual courses offered over the thirteen-year period of data collection. Therefore, we are not able to pinpoint specific practices that may lead to grade penalties, grade bonuses, or gender inequities at our institution. Instead, we assume that, like the courses currently offered, most of these courses focus on teaching in a traditional, lecture-based, and exam-reliant format.

Finally, this research is based at a primarily white, large, public university. While our results are likely to generalize to similar institutions [152, 235], we do not know what patterns of grade anomalies exist at smaller liberal arts colleges, minority-serving institutions, or community colleges in the US. Also, conducting research at a diverse range of institutions in different countries, as well as a focus on how grade anomaly affects students from a variety of underrepresented groups, will help us more fully understand how grade anomalies differ for a range of students.

Table 37: Course Requirements by Major R designates a required course, O designates a course that can be taken for elective credit in the major, and no letter designates a course that does not fulfill any credits for the major. The following terms are abbriviated: Computational (Comp), Biology (Bio), Ecology and Evolution (E&E), Rehabilitation (Rehab), Calculus (Calc), Chemistry (Chem), Genetics (Gen), Organic Chemitry (Organic), Human Physiology (HP), and Biochemistry (BC).

		Bio		Chem			Organic				Physics	
Major	Calc 1	1	2	1	2	Gen	1	2	HP	BC	1	2
Biology	R	R	R	R	R	R	R	R	O	R	R	R
Comp Bio	R	R	R	R	R	R	R			R	O	
E&E	R	R	R	R	R	R	R	R		R	R	R
Microbiology	R	R	R	R	R	R	R	R	O	R	R	R
Molecular Bio	R	R	R	R	R	R	R	R			R	R
Neuroscience	R	R	R	R	R		R	R	R	R	R	R
Pharmacy	R	R	R	R	R		R	R		R	O	O
Rehab Science		R		R					O		R	

Table 38: List of courses studied, the department that offers them, and the percentage of students in our sample who take each course in a given year. For example, 61% of students take calculus during their first year of university, and 11% of students take calculus during their second year. The year in which students take the course most often has its percentage of students in bold.

Course	Department	1^{st}	2^{ed}	3^{ed}	4^{th}	$\geq 5^{th}$
Calculus 1	Mathematics	**61**	22	9	6	2
Biology 1	Biology	**81**	14	3	1	1
Biology 2	Biology	**55**	34	7	2	2
Chemistry 1	Chemistry	**86**	10	2	1	1
Chemistry 2	Chemistry	**64**	28	4	3	1
Genetics	Biology	6	**44**	31	13	6
Organic Chemistry 1	Chemistry	7	**72**	15	4	2
Organic Chemistry 2	Chemistry	3	**50**	32	9	6
Physics 1	Physics	21	**37**	30	7	5
Human Physiology	Biology	2	18	**56**	19	5
Biochemistry	Biology	1	4	**48**	35	12
Physics 2	Physics	6	26	**60**	5	3

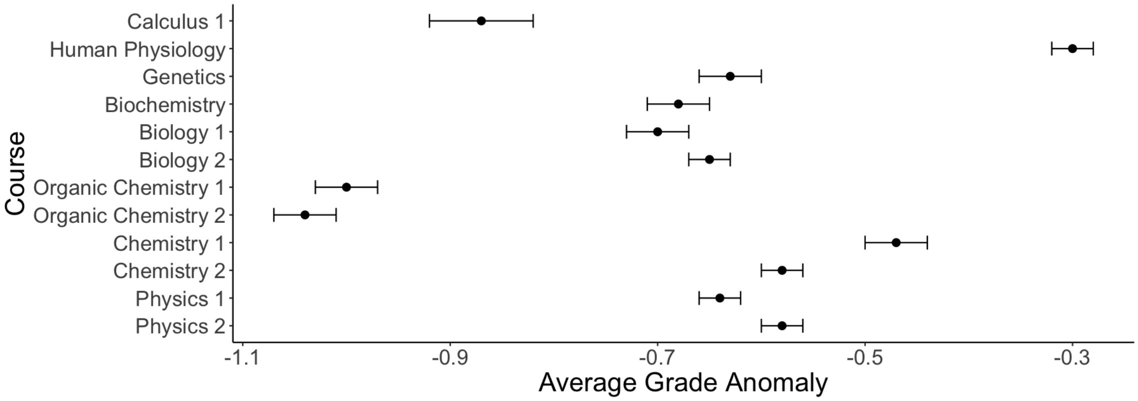


Figure 14: Average grade anomaly (AGA) of all students by course. Ranges represent standard error of the mean.

Table 39: Mean and standard deviation (SD) of average grade anomalies (AGA) and grades, as well as number of students (N) for each course of interest.

			AGA		Grade	
Course	Department	N	Mean	SD	Mean	SD
Calculus 1	Mathematics	1018	-0.87	1.53	2.43	1.22
Human Physiology	Biology	1010	-0.30	0.78	3.11	0.93
Genetics	Biology	775	-0.63	0.89	2.72	1.05
Biochemistry	Biology	886	-0.68	0.89	2.73	1.06
Biology 1	Biology	1740	-0.70	1.10	2.65	1.03
Biology 2	Biology	1630	-0.65	0.78	2.69	0.92
Organic Chemistry 1	Chemistry	1614	-1.00	1.08	2.40	1.15
Organic Chemistry 2	Chemistry	1135	-1.04	1.01	2.35	1.21
Chemistry 1	Chemistry	1746	-0.47	1.08	2.89	0.87
Chemistry 2	Chemistry	1673	-0.58	0.83	2.79	0.96
Physics 1	Physics	2685	-0.64	0.95	2.60	1.08
Physics 2	Physics	1350	-0.58	0.77	2.79	0.97

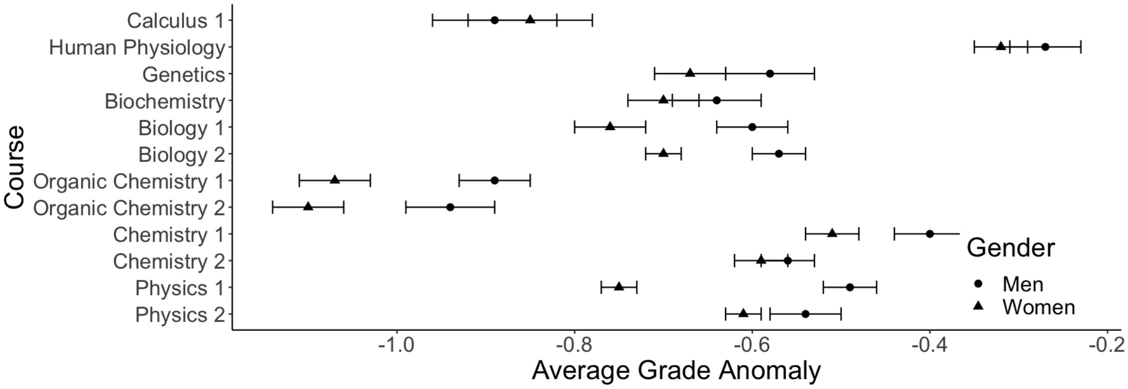


Figure 15: Comparison of average grade anomaly (AGA) between men and women for each course of interest. Ranges represent standard error of the mean.

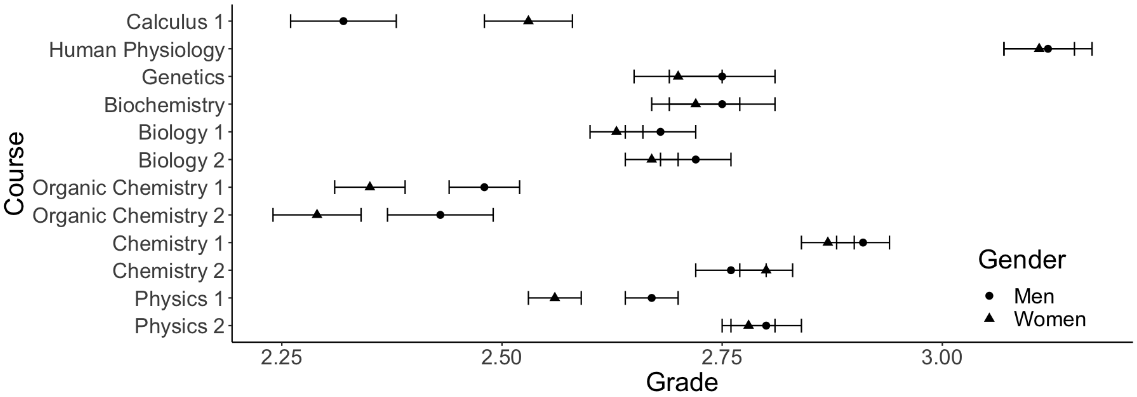


Figure 16: Comparison of average grades between men and women for each course of interest. Ranges represent standard error of the mean.

Table 40: Average grade anomalies (AGAs), grades, and between-gender effect sizes for each course of interest. Cohen's d is positive if men had higher grades or AGAs than women in a course. A bold Cohen's d signifies that a t-test showed significant differences between men and women. The following abbreviations are used: Human Physiology (Hum. Phy.), Biochemistry (Biochem.), Organic Chemistry (Organic), and Chemistry (Chem.). $^{\gamma} = p < 0.05$, $^{\beta} = p < 0.01$, and $^{\alpha} = p < 0.001$.

	Women					Men						
		AGA		Grade			AGA		Grade		Cohen's d	
Course	N	Mean	SD	Mean	SD	N	Mean	SD	Mean	SD	AGA	Grade
Calculus 1	527	-0.85	1.55	2.53	1.17	491	-0.89	1.52	2.32	1.26	-0.03	**-0.18**$^{\beta}$
Hum. Phy.	626	-0.27	0.80	3.11	0.94	384	-0.32	0.76	3.12	0.92	0.07	0.02
Genetics	463	-0.67	0.92	2.70	1.08	312	-0.58	0.84	2.75	0.99	0.10	0.05
Biochem.	508	-0.70	0.88	2.72	1.04	378	-0.64	0.91	2.75	1.09	0.07	0.02
Biology 1	1078	-0.76	1.16	2.63	1.03	662	-0.60	1.00	2.68	1.05	**0.14**$^{\beta}$	0.04
Biology 2	993	-0.70	0.77	2.67	0.90	637	-0.57	0.78	2.72	0.94	**0.16**$^{\beta}$	0.05
Organic 1	983	-1.07	1.12	2.35	1.16	631	-0.89	1.02	2.48	1.11	**0.16**$^{\beta}$	**0.11**$^{\gamma}$
Organic 2	673	-1.10	1.00	2.29	1.19	462	-0.94	1.02	2.43	1.22	**0.16**$^{\beta}$	0.11
Chem. 1	1046	-0.51	1.11	2.87	0.87	700	-0.40	1.04	2.91	0.88	**0.10**$^{\gamma}$	0.04
Chem. 2	986	-0.59	0.80	2.80	0.92	687	-0.56	0.88	2.76	1.02	0.03	-0.05
Physics 1	1572	-0.75	0.94	2.56	1.04	1113	-0.49	0.95	2.67	1.14	**0.28**$^{\alpha}$	**0.11**$^{\beta}$
Physics 2	776	-0.61	0.68	2.78	0.89	574	-0.54	0.86	2.80	1.06	0.09	0.02

11.0 Grades and grade anomalies before, during, and after remote COVID-19 instruction for first-year engineering majors: Overall trends and gender inequities

11.1 Introduction

Remote teaching due to the COVID-19 pandemic has inspired research assessing the differences between online and in-person courses regarding student learning outcomes and classroom equity [142–145]. There are mixed findings regarding the effect of online instruction on student learning [142, 143]. In this study, we explore overall trends in both grades and grade anomalies before, during, and after the period of remote instruction due to COVID-19 in courses for first-year engineering students in a large public university.

We define grade anomaly as the difference between a student's grade in a course of interest and their grade point average (GPA) in all other classes up to that point. The mean of this statistic for all students who took a course is the average grade anomaly (AGA). We divide average grade anomalies into "bonuses" and "penalties". A course in which students on average earn a lower grade than usual has an AGA with grade penalty, while a course in which students on average earn a higher grade than usual has an AGA with grade bonus.

Within our framework, we posit that grade anomaly may allow us to track, through institutional grade data, an important measure of how courses may affect students' academic self-concept. Academic self-concept is a relatively stable measure of a students' perceived ability to succeed in the academic sphere, and is based on grades and outside feedback (e.g., from parents, peers, and instructors) [78–80, 233, 236]. Grades inform academic self-concept as both an external ("How good at math am I compared to other students?") and internal ("How good am I at math compared to English?") frame of reference [79, 80, 233]. We also note that, while academic self-concept is generally quite stable, it can change quite quickly during periods of transition (such as the transition from high school to university) [233]. Grade penalties in STEM courses during the first two semesters (but not later semesters) of university were negatively correlated with completing a STEM degree, even when controlling for gender, race, high school preparation, and college performance [239]. These findings hint at the importance of monitoring and minimizing grade penalties in students' first few semesters.

Our framework uses grade penalty as a central construct instead of grade because students' academic self-concept is often based on comparisons, not absolute grades [78]. Students may compare their grades across courses to determine which disciplines they excel at or struggle with [78]. Additionally, students tend to have a fairly fixed view of what "kind" of student they are, e.g., students may endorse the idea that "If I get As, I must be an A kind of person. If I get a C, I am a C kind of person" [53]. Grade anomalies may challenge or reinforce students' ideas about what kind of student they are, and if they are capable of succeeding in their chosen major. Many students who leave STEM majors explicitly cite lower grades than they are used to as a reason for doing so [53,54]. Grade penalties are more common and extreme in STEM disciplines than in humanities or social science departments [53,152,234,235].

Additionally, we aim to investigate gender differences in grades and AGAs. When women leave STEM disciplines, they often do so with higher grades than men who remain in the program [39,54,59]. Women are more underrepresented in engineering than in many other STEM disciplines [4,7,11,12,41,43,45,65], so focusing on retention is important for this field. If women are leaving engineering programs with grades that meet or exceed minimum requirements [53, 54], it is likely that many students who would succeed in engineering careers will pursue other professional paths.

Broadly, in this research we aim to understand differences in students' grade anomalies before, during, and after the period of remote instruction due to COVID-19 with a particular focus on gender differences in grades and grade anomalies. This will build on previous work which observed grade anomalies at this same institution for over ten years pre-COVID [14, 15, 19]. We focus on first-year engineering majors and aim to answer the following research questions regarding grade anomalies:

RQ1. Do grades or grade anomalies differ between before, during, and after the period of remote COVID teaching?

RQ2. Are there gender differences in grades or grade anomalies, and do they differ between the periods before, during, and after COVID-19 remote instruction?

11.2 Methodology

11.2.1 Participants

Participants in this study were enrolled in engineering majors at a large, public, and urban institution. Grade data were collected over four years. We divide these semesters into three groups, which are described in Table 41. We excluded courses that were taken during the summer semester. We excluded summer courses because they are not a typical representation of courses at our institution. For example, many summer students do not primarily attend our institution, but are local students visiting home for the summer. In addition the class sizes are an order of magnitude smaller than those in the Fall and Spring semesters.

Table 41: Labels for each time period studied

Label	Period
Pre-Remote	Four semesters of in-person instruction before the COVID-19 pandemic, excluding Spring 2020
Remote	Two semesters of remote instruction due to the COVID-19 pandemic, excluding Spring 2020
Post-Remote	Two semesters of in-person instruction after the return to in-person classes

This left us with 5,807 pre-remote, 2,775 remote, and 4,065 post-remote instances of an enrollment in a course. For example, a student who takes four courses in one semester and three in the next semester has seven instances of enrollment. Demographic information for the student sample can be found in Table 42. De-identified demographic data were provided through university records.

11.2.2 Course Selection

At this institution, there is a standardized curriculum for first-year engineering majors. All of these courses were included in this research, with an exception of two pass/fail seminars which do not count towards a students' grade point average. This included a total of ten courses, which are described in Table 43. Students typically took Physics 1, Chemistry 1, Calculus 1, Engineering Analysis, and Composition Seminar during their first Fall semester. Students typically took Physics 2, Chemistry 2, Calculus 2, Engineering Computing, and Engineering Communication during their first Spring semester. Engineering Communication was not offered until Spring of 2020, so there is no Pre-Remote data for this course.

Table 42: Demographic information for study participants. Several survey options for ethnicity were excluded because they each made up less than 0.5% of the sample. These groups are Indigenous American, Pacific Islander, Not Specified, and Other. Unknown indicates that a student did not submit a response to the item, while Not Specified indicated that they chose the option "I prefer not to specify".

	Sex		Race/Ethnicity					
Group	Female	Male	Asian	Black	Latine	Multiracial	White	Unknown
Pre-Remote	36%	64%	11%	4%	6%	5%	74%	1%
Remote	36%	64%	9%	5%	6%	6%	72%	1%
Post-Remote	31%	69%	17%	4%	6%	5%	65%	2%

11.2.3 Measures

11.2.3.1 Course Grade

Course grades were based on the 0-4 scale used at our university, and a conversion of letter grades to GPA points can be seen in Table 44. We are unable to report grading schemes of each instructor, type of course (i.e., traditional lectures or active learning), or any other detailed course-level information due to the large number of courses sampled.

11.2.3.2 Grade Anomaly

GA was found by first finding each student's grade point average excluding the course of interest (GPA_{exc}). This was done by using the equation

$$GPA_{exc} = \frac{GPA_c \times Units_c - Grade \times Units}{Units_c - Units} \tag{11}$$

where GPA_c is the student's cumulative GPA, $Units_c$ is the cumulative number of units the student has taken (also called credit hours), $Grade$ is the grade the student received in an individual course, and $Units$ is the number of units associated with an individual course. After finding GPA_{exc} we can calculate grade anomaly (GA) by finding the difference between a student's GPA_{exc} and the grade received in that class:

$$GA = Grade - GPA_{exc}. \tag{12}$$

A negative GA corresponds to a course grade lower than a students' GPA in other classes and we call this a "grade penalty". A positive GA corresponds to a course grade higher than a students' GPA in other classes and we call this a "grade bonus". Average grade anomaly (AGA) is the mean of students' grade anomalies (GA) for each course, and is the metric by which we compare courses.

11.2.3.3 Analysis

To characterize both average grade anomaly (AGA) and grades, we found the sample size, mean, standard deviation, and standard error of each measurement for each course of interest. We calculated these statistics for women and men separately, and then for all students combined. We also compared the effect size of gender on both grade and grade anomaly, using Cohen's d to describe the size of the mean differences and unpaired t-tests to evaluate the statistical robustness of the differences. Cohen's d is calculated as follows:

$$d = \frac{\mu_1 - \mu_2}{\sqrt{\sigma_1^2 - \sigma_2^2}} \tag{13}$$

where μ_1 and μ_2 are the means of the two groups, and σ_1 and σ_2 are the standard deviations [108]. Cohen's d is considered small if $d \sim 0.2$, medium if $d \sim 0.5$, and large if $d \sim 0.8$ [140]. We used a significance level of 0.05 in the t-tests as a balance between Type I (falsely rejecting a null hypothesis) and Type II (falsely accepting a null hypothesis) errors [108]. All analysis was conducted using R [174], using the package plotrix [245] for descriptive statistics, lsr [246] for effect sizes, and ggplot2 [247] to create plots.

11.3 Results

11.3.1 Chemistry Courses

Chemistry courses had the lowest grades of the courses studied during the pre-remote, remote, and post-remote periods, which can be seen in Figures 17 and 18 as well as Tables 69 and 70 in Appendix J. For Chemistry 1 and 2, grades as well as grade anomalies were similar before and during remote instruction. However, during post-remote instruction, average course grades (see Figure 17) and the magnitude of the grade penalty (see Figure 18) increased. Before and during remote instruction the average grade for Chemistry 1 was between a C+ and B− (2.48 for both), which dropped to a C+ (2.24) for post-remote instruction. Before and during remote instruction the average grade for Chemistry 2 was also between a C+ and B− (2.36 and 2.49, respectively), which also dropped to a C+ (2.24) for post-remote instruction.

Chemistry courses also had the largest grade penalties during the pre-remote,

remote periods, which can be seen in Figure 18. Excluding Calculus 1, they also had the largest grade penalties in the post-remote period. Chemistry 1 had slightly larger grade penalties than Chemistry 2. Students taking Chemistry 1 could expect a grade approximately three-fourths of a letter grade lower than their overall GPA before and during remote instruction, and a full letter grade after remote instruction. Students could generally expect a grade three-fourths of a letter grade lower than their other courses for all time periods studied. Table 45 shows that neither chemistry course had any statistically significant difference between men's and women's grades or average grade anomalies.

11.3.2 Engineering Courses

Generally, courses offered by the Engineering School had the highest grades of all STEM courses, and were the only STEM courses that had a grade bonus (or a grade anomaly of almost 0) rather than grade penalty which can be seen in Figures 17 and 18 as well as Tables 69 and 70 in Appendix J. For Engineering Analysis, average grades increased slightly from pre-remote to remote instruction and decreased slightly from remote to post-remote instruction, though the average grade remained between a B+ and A− during all three time periods. Figure 18 reveals that, on average, students tended to have a grade bonus of almost half a letter grade. Generally, courses offered by the Engineering School had the highest grades of all STEM courses, and were the only STEM courses that had a grade bonus (or a grade anomaly of almost 0) rather than grade penalty which can be seen in Figures 17 and 18 as well as Tables 69 and 70 in Appendix J. For Engineering Analysis, average grades increased slightly from pre-remote to remote instruction and decreased slightly from remote to post-remote instruction, though the average grade remained between a B+ and A−

during all three time periods. Figure 18 reveals that, on average, students tended to have a grade bonus of almost half a letter grade. Generally, courses offered by the Engineering School had the highest grades of all STEM courses, and were the only STEM courses that had a grade bonus (or a grade anomaly of almost 0) rather than grade penalty which can be seen in Figures 17 and 18 as well as Tables 69 and 70 in Appendix J. For Engineering Analysis, average grades increased slightly from pre-remote to remote instruction and decreased slightly from remote to post-remote instruction, though the average grade remained between a B+ and A− during all three time periods. Figure 18 reveals that, on average, students tended to have a grade bonus of almost half a letter grade.

For Engineering Computing, average grades dropped slightly from pre-remote to remote instruction and again from remote to post-remote instruction (see Figure 18 and Table 70). However, the average grade remained between a B and B+ throughout. Before remote instruction, students tended to have a slightly higher grade in Engineering Computing than their average, but during and after remote instruction there was no grade anomaly in this course.

There were generally no statistically significant grade or average grade anomaly differences between men and woman in these courses, which can be seen in Table 46. One exception was average grade anomaly during post-remote courses, in which men had a small grade bonus and women had a small grade penalty.

11.3.3 English Courses

Courses offered by the English Department were the only non-STEM courses included in this study, and Figures 17 and 18 show that they also tended to have the highest grades and largest grade bonuses of all the courses included in this research.

The average grade for Composition Seminar was between a B+ and A− throughout, though the average course grade was slightly lower during post-remote instruction than pre-remote or remote instruction. During all three periods, students on average had a grade half of a letter grade higher in Composition Seminar then in their other courses (see Figure 18).

There is no pre-remote instruction data for Engineering Communication because the class did not exist yet. However, Figure 17 shows that the average grade during remote instruction was the highest of any course studied: an A− (3.80). During post-remote instruction, the average grade decreased slightly to 3.60. Composition Seminar had the largest graded bonuses of all the courses, and students generally had almost three-fourths of a letter grade higher in this course than in their other courses during remote and post-remote instruction (see Figure 18).

In Composition Seminar, Table 47 shows that there were statistically significant gender differences in both grades and grade anomalies. Before and during remote instruction, women tended to have higher grades and larger grade bonuses than men, and after remote instruction women had larger average grade bonuses (but not grades) than men. There were no statistically significant grade or average grade anomaly differences between men and women for Engineering Communication either during or after remote instruction which can be seen in Table 47.

11.3.4 Mathematics Courses

Unlike courses offered by other departments, the courses in the Mathematics department, Calculus 1 and 2, did not follow similar trends. Figures 17 and **??** reveal that the average grade in Calculus 1 went from approximately a B− during the pre-remote and remote periods, and dropped to a C+ during the post-remote period. Though the average letter grade was the same during the pre-remote and remote periods, the average grade in Calculus 1 decreased from pre-remote to remote instruction. ConcerningLY , Calculus 1 was the only course in which the average grade consistently decreased from pre-remote to remote to post-remote courses. On the other hand, the average Calculus 2 grade was between a C+ and B− during the pre-remote period, rose to a B during remote teaching, and fell back to a 2.30 during the post-remote period. This was a common trend among the overall set of courses (see Figure 17): grades were similar during the pre- and post-remote periods, but slighly higher during the remote period.

Regarding AGAs, Figure 18 shows that Calculus 1 had a comparatively small grade penalty compared to other courses during the pre-remote period. However, the AGA for Calculus 1 increased in magnitude for each period. In fact, Calculus 1 had the largest grade penalties aside from the Chemistry courses during remote and post-remote instruction. Calculus 2 consistently had AGAs that were not particularly high or low compared to other courses studied. The average grade penalty in Calculus 2 was identical during the pre- and post-remote periods, but was smaller during the remote period.

There were no statistically significant gendered grade differences in either Mathematics course during any period studied, which can be seen in Table 48. There was a gender difference in AGA in Calculus 1 during the post remote period, with women having larger average grade penalties than men. Aside from this, there were no gendered differences in AGAs.

11.3.5 Physics Courses

Figure 17 and Table 69 show that Physics 1 letter grades increased from pre to during, but then decreased again during post. In Physics 1, grades went from a B− during pre-remote courses to a B during remote instruction to between a C+ and B− during Post-Remote instruction. Physics 2 letter grades increased from a B− to a B from pre-remote to remote teaching. However, instead of decreasing again during post-remote instruction, the grades remained consistent, and the average grade during post-remote instruction was also a B−. Physics 1 and 2 average grade penalties followed similar trends. Both had similar AGAs during pre- and post-remote instruction, and had smaller AGAs during remote instruction.

Courses offered by the physics department tended to have more gender differences in both grades and AGAs than courses offered by other departments, which can be seen in Table 49. During pre-remote instruction both Physics 1 and 2 had gendered grade differences. In both cases, men on average had higher grades than women, with a small effect size for both courses ($d \sim 0.2$). There were also gendered grade differences in Physics 1 grades during post-remote instruction which were similar in magnitude to pre-remote gender differences. There were gender differences in AGAs for both Physics courses during almost all periods, as shown in Table 49. Physics 1 had a small gender differences ($d \sim 0.2$) during pre-remote and remote courses, and had medium gender differences ($d \sim 0.5$) during post-remote instruction. Physics 2 had medium ($d \sim 0.5$) AGA gender differences during pre-remote instruction and small-to-medium ($d \sim 0.2$ to 0.5) gender differences during post-remote instruction.

11.4 Discussion

11.4.1 Do grades or grade anomalies differ between before COVID, during remote COVID teaching, and after remote COVID teaching?

Grades are important to students for a variety of reasons such as continuing their major, scholarship requirements, graduate school or professional school admissions, and career goals. In general, grades were higher during remote instruction than they were during pre-remote instruction, and then decreased after remote instruction. During remote instruction, grades tended to be a fraction of a letter grade higher than during pre- or post-remote instruction (for example, the mean Calculus 2 grade was C+ during pre-remote instruction, B− during remote instruction, and a C+ during post-remote instruction). These increases in grades may be due to a range of factors. For example, grading schemes and assessment types may have been changed, or instructors may have been more flexible than during pre-remote classes [258].

Broadly, grades were higher during remote instruction and were lower again during post-remote instruction but there were some courses that did not follow this trend. Two of those courses will not be discussed here because they had higher grades compared to most courses in this study. On the other hand, Chemistry 1, Chemistry 2, and Calculus 1 had concerning trends in grades. Namely, Chemistry 1 and 2 had the lowest overall grades of the courses studied during all periods studied: both had a C average post-remote grade. Calculus 1 had the largest decrease in average grade of any course over time - from pre- to post-remote, the average grade dropped from a 2.89 to a 2.30 on a 4-point scale.

AGAs, unlike grades, do not have a direct effect on students' outcomes such as scholarships and graduate admissions. A student with an A average who receives a B in a class has the same grade anomaly as a student with a B average who receives a C in the class. Here, we use the idea of academic self concept from Situated Expectancy Value Theory to frame how students may think about grade anomalies [78]. AGAs may challenge a student's idea about what kind of student they are (i.e. an "A" student or a "C" student) [53]. In particular, students may compare their grades across courses to determine which disciplines they excel at or struggle with [78].

Our results show that there are grade penalties in all Chemistry, Math, and Physics courses studied, while there were either grade bonuses or no grade anomaly in the Engineering and English Composition courses. Other studies that focus on AGA find that science and math courses have large grade penalties, while humanities courses have grade bonuses [53, 152, 234, 235]. This aligns with our findings except that engineering courses do not have grade penalties.

Generally, AGAs had a smaller magnitude during remote instruction than pre- or post-remote instruction. That is, generally, students' grades were more consistent during remote instruction, so that most classes deviated less from a students GPA during remote instruction. This was not true for Calculus 1, Chemistry 1, or Chemistry 2. The Chemistry courses did not have increased average grades during remote instruction as many other courses did, while Calculus 1 actually had lower grades during remote than pre-remote instruction. Throughout the study, Chemistry 1 and 2 had the largest grade penalties.

We hypothesize that smaller grade anomalies may result in students being less concerned that they can succeed in their discipline, and may rely more on other factors (such as interest) to make decisions regarding major and career choice.

Grade penalties are more common and larger in STEM disciplines than in social sciences or humanities [53, 152, 234, 235], but our findings show that there are significant variations in AGAs even among STEM courses. For example, Chemistry courses tended to have large grade penalties, Engineering courses tended to have grade bonuses, and Physics and Mathematics courses tended to have AGAs in the middle. Thus, AGAs are not a simple issue of STEM courses having larger AGAs than non-STEM courses. Instructors and departments with comparatively lower grades and larger AGAs than others may benefit from pedagogies implemented by other STEM departments and instructors at their institution.

There are likely to be a range of potential factors contributing to differences in grades and AGAs over time. Though there is a possibility that some students are cheating, cheating on exams seems to have only small increases in the USA during the pandemic, though the effect may be larger in other regions [259]. There is also research that suggests that there are specific factors that could lead to increased grades during remote instruction. For example, because there were were more low-stakes assessments during remote instruction, students may be more likely to engage in spaced practice instead of “cramming” for assessments during remote learning [260]. One study showed that students had higher grades during COVID-19 remote instruction even on identical assessments that were also given online pre-pandemic [260]. One study that focused on quantum mechanics (an upper-level physics course) found that implementing low-stakes formative assessments instead of exams did not lead to lower scores on course post-test (which only contributed a small amount to the students’ final grade) [143]. Though these studies do not specify any specific reason that there may be differences in grades during remote versus in-person classes, they do suggest that increases in grades do not necessarily correlate with lowered academic standards or cheating.

11.4.2 RQ2. Are there gender differences in grades or grade anomalies, and do they differ between before COVID, during remote COVID teaching, and after remote COVID teaching?

During pre-remote instruction, three courses had statistically significant gender differences. For Physics 1 and 2, men had higher grades then women. For Composition Seminar, women had higher grades then men. During remote instruction, only composition seminar had statistically significant grade differences. Again women had higher grades then men. Finally, during post-remote instruction, men had higher grades than women in Physics 1. Physics 1 gendered grade differences were very similar between pre-remote and post-remote courses.

There were more instances of statistically significant AGA differences than grade differences between men and women. During pre-remote instruction, both Physics courses and Composition Seminar had gender differences. Compared to women, men had smaller grade penalties in Physics 1 and 2 as well as smaller grade bonuses in Composition Seminar. During remote instruction, there were similar trends. Men had smaller grade penalties in Physics 1 and smaller grade bonuses in Composition Seminar, and the effect size of these differences did not change substantially between pre-remote and remote instruction.

The post-remote period had more AGA differences by gender than the other periods. Men had smaller grade penalties than women in Physics 1, Physics 2, and Calculus 1. Women had larger grade bonuses then men in Composition Seminar. In Engineering Computing, men had a small grade bonus and women had a small grade penalty. Broadly, we note that there are more gender differences in AGA than in grades, and that Physics and Composition Seminar had more gender differences in grades and AGAs than other courses. However, because students tend to have grade bonuses in Composition Seminar, we are less worried about this course.

For women in engineering majors, a large grade anomaly in their first Physics course at university may be particularly concerning, and potentially lead them to believe that they do not "have what it takes" to succeed in their major. Women often report worrying more then men that they do not understand the material even if they receive A's, B's, or C's (which are grades that allow students to continue in most programs) [53]. This trend has been found to be particularly strong among high-achieving women [53].

We hypothesize that women may be more likely to have a low academic-self-concept than men at similar performance levels. Prior work has theorized that men are more likely to separate their grades and sense of academic self-concept [53,54,237]. Academic self-concept is formed through grades and feedback from outsiders. Women are generally less likely to receive recognition from instructors [56,130,208], so women may rely more than men on grade information to develop their academic self-concept [53, 54, 237]. Women also tend to earn higher grades than men who have the same standardized test scores [53,203], so they may be more accustomed to higher grades. As a result, they may have more concern about grades that are lower than what they are accustomed to, especially during the transition from high school to university.

AGAs and raw grade data do not always reveal the same trends: there are many more gender differences in AGA than in grades in the findings presented here. This trend reveals how AGA may be a useful measure. For example, an instructor may not see any gender differences in grades, which is one important indicator of gender equity. However, if they do not know the gender differences in AGA, an instructor or department may not recognize how those grades may be perceived by women and men in their classes. Understanding both grades and AGA differences may allow instructors to understand classroom-level inequities better.

11.5 Conclusion, Limitations, and Future Research

In this work we found that courses offered by the Engineering and English departments tended to have grade bonuses while courses offered by the Physics, Mathematics, and Chemistry departments tended to have grade penalties. Generally, grades were higher and grade penalties were smaller during remote instruction compared to pre-remote instruction. During post-remote instruction, grades were lower and grade penalties were larger than during remote instruction. Further, there were more gender differences in both grades and grade penalties (favoring men) during post-remote teaching than for pre-remote or remote teaching.

These results are very important because they provide evidence that courses in STEM departments tend to have grade penalties, and that these penalties tended to decrease during remote instruction. Additionally, AGA may also act as a useful measure of academic self-concept that is easy for institutions to access.

Although we have evidence of grade penalties in the studied courses as well as gendered grade anomaly differences, we did not have access to syllabi or other information about individual courses offered over the period of data collection. Therefore, we are not able to pinpoint specific practices that may lead to grade penalties, grade bonuses, or gender inequities at our institution.

Finally, this research is based at a primarily white, large, public university. While our results may generalize to similar institutions, we do not know what patterns of grade anomalies exist at smaller liberal arts colleges, minority-serving institutions, or community colleges in the US. Additionally, it may also be useful to repeat similar research in other countries, as many countries worldwide were affected differently by the COVID-19 pandemic.

Table 43: Courses engineering majors were required to take during their first year, along with which department/school offered the course and a description of the course. Engineering its own school in the university.

Course Title	Department	Description
Physics 1	Physics	Calculus-based, covered mechanics and waves
Physics 2	Physics	Calculus-based, covered electricity, magnetism, circuits, electromagnetic theory and optics
Chemistry 1	Chemistry	Only for engineering students. Covered stoichiometry, the properties of solids, liquids and gases, thermochemistry and the electronic structure of atoms and molecules.
Chemistry 2	Chemistry	Only for engineering students. Covered solutions, thermodynamics, kinetics, chemical equilibrium, coordination chemistry, redox reactions and nuclear chemistry.
Calculus 1	Mathematics	Covered derivative and integral of functions of one variable and their applications.
Calculus 2	Mathematics	Covered calculus of transcendental functions, techniques of integration, series of numbers and functions, polar coordinates, and conic sections
Eng Analysis	Engineering	Covered an introduction to Excel and an introduction to design and entrepreneurship.
Eng Computing	Engineering	Covered basic programming skills using MATLAB and C.
Composition Seminar	English	Course in which students wrote about the disciplines, practices, methods, ethics, and education of engineering.
Eng Communication	English	Students researched and wrote about a single topic regarding a current engineering innovation or technology in a conference paper format.

Table 44: Grades and GPA points for this university's grading standards. For most majors, a "C" or above is a passing grade. A C was the minimum grade needed to pass a course at this institution for all majors included.

Grade	A/A+	A−	B+	B	B−	C+	C	C−	D+	D	D−	F
GPA Value	4.00	3.75	3.25	3.00	2.75	2.25	2.00	1.75	1.25	1.00	0.75	0.00

Table 45: Means and standard deviations (SD) of grades and grade anomalies by gender for courses offered by the Chemistry ("Chem") Department before ("Pre"), during ("Rem"), and after ("Post") remote instruction due to COVID-19. Cohen's d is positive if men had higher grades or smaller AGAs than women in a course. $^{\gamma} = p < 0.05$, $^{\beta} = p < 0.01$, and $^{\alpha} = p < 0.001$.

		Women					Men						
			AGA		Grade			AGA		Grade		Cohen's d	
Course	Type	N	Mean	SD	Mean	SD	N	Mean	SD	Mean	SD	AGA	Grade
Chem 1	Pre	234	-0.83	1.17	2.50	1.00	418	-0.77	1.01	2.47	0.99	0.05	-0.03
	Rem	79	-0.74	0.74	2.53	0.91	167	-0.75	0.73	2.46	0.92	-0.01	-0.07
	Post	106	-1.11	1.41	2.23	1.25	257	-0.91	0.96	2.25	1.15	0.18	0.01
Chem 2	Pre	148	-0.74	0.55	2.39	0.77	292	-0.73	0.59	2.34	0.87	0.02	-0.06
	Rem	70	-0.60	0.52	2.63	0.80	135	-0.75	0.63	2.41	0.87	-0.25	0.26
	Post	82	-0.81	0.62	2.27	0.94	205	-0.83	0.69	2.24	1.05	-0.04	-0.03

Table 46: Means and standard deviations (SD) of grades and grade anomalies by gender for courses offered by the Engineering School before (“Pre”), during (“Rem”), and after (“Post”) remote instruction due to COVID-19. Engineering Analysis is abbreviated as “Analysis”, and Engineering Computing is abbreviated as “Comp”. Cohen’s d is positive if men had higher grades or smaller AGAs than women in a course. $^{\gamma} = p < 0.05$, $^{\beta} = p < 0.01$, and $^{\alpha} = p < 0.001$.

		Women					Men						
			AGA		Grade			AGA		Grade		Cohen’s d	
Course	Type	N	Mean	SD	Mean	SD	N	Mean	SD	Mean	SD	AGA	Grade
Analysis	Pre	320	0.36	0.53	3.43	0.57	531	0.36	0.60	3.37	0.6	0.01	0.10
	Rem	132	0.42	0.48	3.53	0.65	212	0.45	0.45	3.47	0.64	0.08	-0.09
	Post	156	0.40	0.72	3.28	0.82	322	0.45	0.62	3.33	0.76	0.08	0.07
Comp	Pre	231	0.10	0.79	3.15	0.92	439	0.22	0.73	3.23	0.80	0.15	0.10
	Rem	103	-0.06	0.57	3.10	0.78	187	0.03	0.62	3.15	0.82	0.15	0.06
	Post	126	-0.14	0.74	2.89	0.98	289	0.08	0.71	3.06	0.94	0.29^{γ}	0.17

Table 47: Means and standard deviations (SD) of grades and grade anomalies by gender for courses offered by the English Department before ("Pre"), during ("Rem"), and after ("Post") remote instruction due to COVID-19. . Composition Seminar is abbreviated as "Seminar", and Engineering Communication is abbreviated as "Comm". Cohen's d is positive if men had higher grades or smaller AGAs than women in a course. $^{\gamma} = p < 0.05$, $^{\beta} = p < 0.01$, and $^{\alpha} = p < 0.001$.

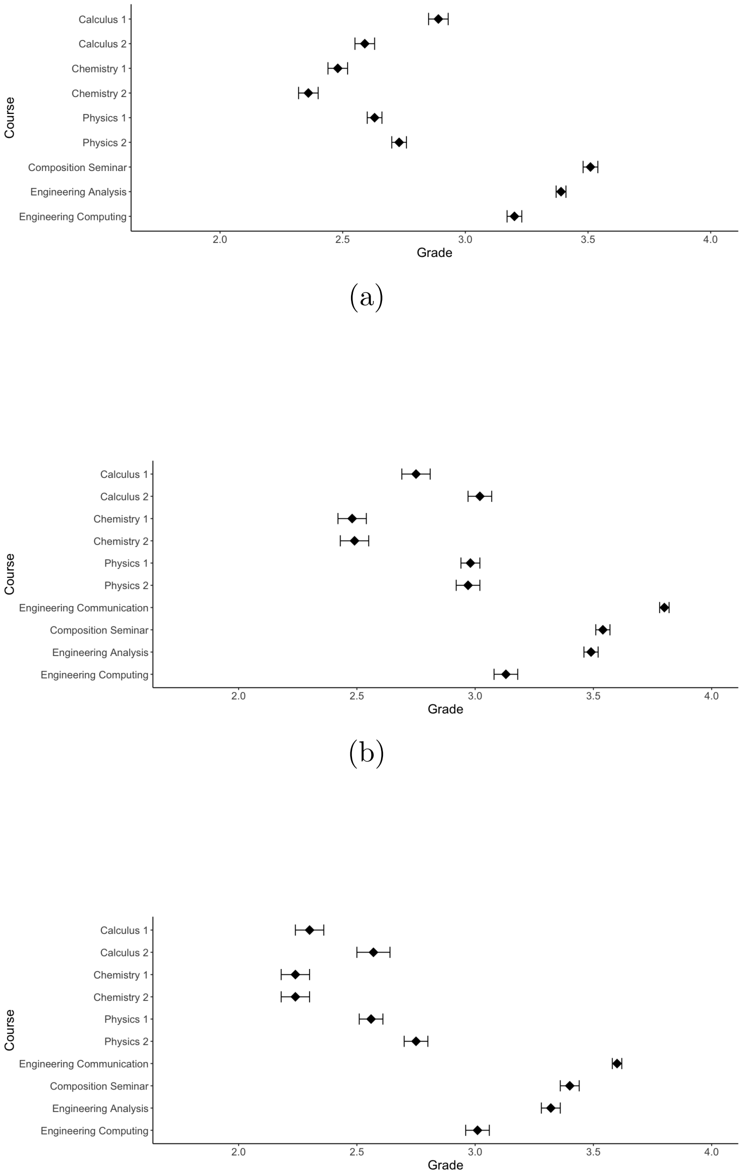
		Women					Men						
			AGA		Grade			AGA		Grade		Cohen's d	
Course	Type	N	Mean	SD	Mean	SD	N	Mean	SD	Mean	SD	AGA	Grade
Seminar	Pre	161	0.59	0.79	3.68	0.58	281	0.34	0.74	3.42	0.77	-0.33$^{\beta}$	-0.36$^{\alpha}$
	Rem	138	0.57	0.60	3.67	0.53	223	0.43	0.53	3.45	0.55	-0.25$^{\gamma}$	-0.39$^{\alpha}$
	Post	164	0.62	0.81	3.48	0.81	347	0.43	0.90	3.36	0.83	-0.21$^{\gamma}$	-0.14
Comm	Rem	104	0.75	0.54	3.83	0.30	191	0.73	0.60	3.78	0.37	-0.03	-0.17
	Post	126	0.67	0.63	3.62	0.43	284	0.64	0.67	3.59	0.43	-0.05	-0.08

Table 48: Means and standard deviations (SD) of grades and grade anomalies by gender for courses offered by the Mathematics Department before (“Pre”), during (“Rem”), and after (“Post”) remote instruction due to COVID-19. Calculus is abbreviated as “Calc”. Cohen’s d is positive if men had higher grades or smaller AGAs than women in a course. $^{\gamma} = p < 0.05$, $^{\beta} = p < 0.01$, and $^{\alpha} = p < 0.001$.

		Women					Men						
			AGA		Grade			AGA		Grade		Cohen’s d	
Course	Type	N	Mean	SD	Mean	SD	N	Mean	SD	Mean	SD	AGA	Grade
Calc 1	Pre	216	-0.22	1.45	2.97	0.99	365	-0.28	1.43	2.85	1.00	-0.04	-0.12
	Rem	94	-0.56	1.11	2.74	0.82	148	-0.43	1.02	2.76	0.95	0.12	0.01
	Post	155	-1.21	2.31	2.21	1.33	310	-0.68	1.42	2.35	1.22	0.30^{γ}	0.11
Calc 2	Pre	209	-0.55	1.06	2.61	1.16	395	-0.54	0.98	2.57	1.07	-0.01	-0.03
	Rem	78	-0.31	0.51	2.96	0.68	133	-0.15	1.00	3.05	0.78	0.19	0.11
	Post	92	-0.41	0.83	2.71	1.05	231	-0.61	1.05	2.51	1.22	-0.20	-0.17

Table 49: Means and standard deviations (SD) of grades and grade anomalies by gender for courses offered by the Physics Department before ("Pre"), during ("Rem"), and after ("Post") remote instruction due to COVID-19. Physics is abbreviated as "Phys". Cohen's d is positive if men had higher grades or smaller AGAs than women in a course. $^{\gamma} = p < 0.05$, $^{\beta} = p < 0.01$, and $^{\alpha} = p < 0.001$.

		Women					Men						
			AGA		Grade			AGA		Grade		Cohen's d	
Course	Type	N	Mean	SD	Mean	SD	N	Mean	SD	Mean	SD	AGA	Grade
Phys 1	Pre	376	-0.84	0.94	2.49	0.77	573	-0.51	1.03	2.71	0.86	0.33^{α}	0.27^{α}
	Rem	127	-0.32	0.64	2.94	0.71	209	-0.12	0.63	3.00	0.7	0.32^{γ}	0.09
	Post	180	-0.95	1.55	2.36	1.03	351	-0.45	1.05	2.66	1.03	0.40^{α}	0.29^{β}
Phys 2	Pre	198	-0.67	0.55	2.59	0.74	416	-0.39	0.60	2.80	0.82	0.48^{α}	0.27^{β}
	Rem	87	-0.28	0.54	3.04	0.73	157	-0.34	0.69	2.93	0.81	-0.09	-0.15
	Post	88	-0.61	0.50	2.72	0.74	219	-0.42	0.56	2.76	0.81	0.35^{β}	0.05

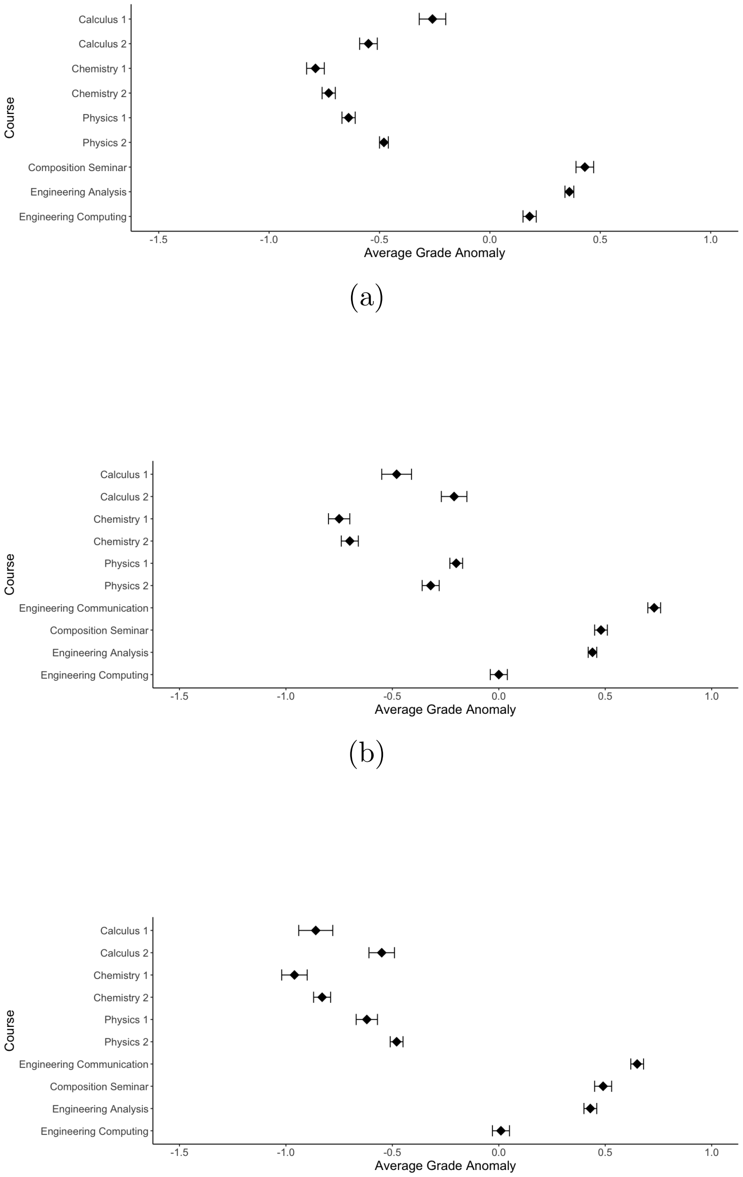


(a)

(b)

(c)

Figure 17: Comparison of student grades for each course of interest for classes before COVID (a), during remote instruction (b), and during Post-Remote instruction (c). Ranges represent standard error of the mean.

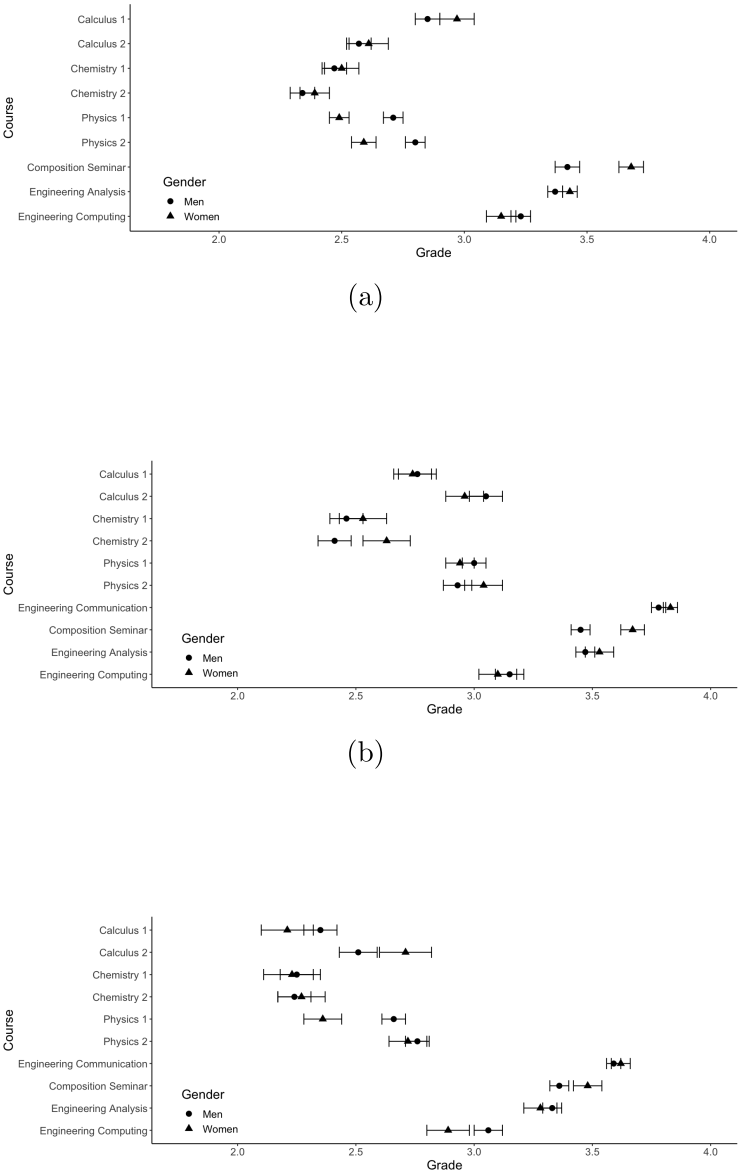


(a)

(b)

(c)

Figure 18: Comparison of student average grade anomalies for each course of interest for classes before COVID (a), during remote instruction (b), and during Post-Remote instruction (c). Ranges represent standard error of the mean.

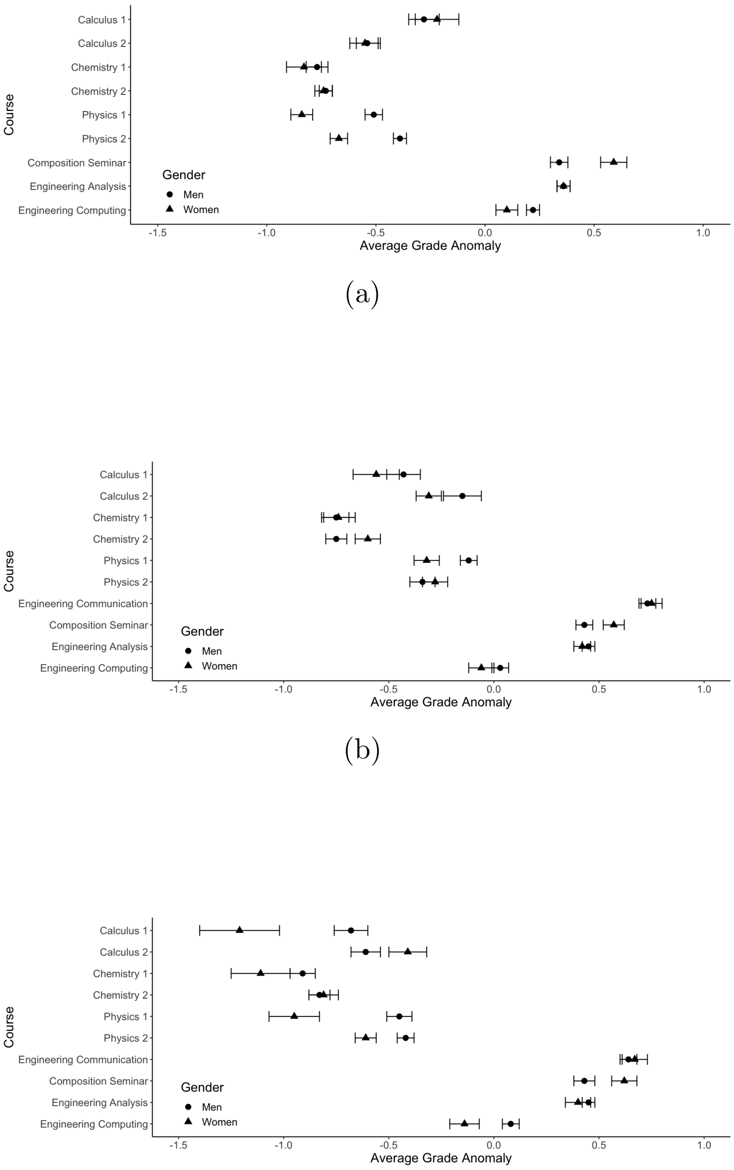


(a)

(b)

(c)

Figure 19: Comparison of student grades for men and women for each course of interest for classes before COVID (a), during remote instruction (b), and during Post-Remote instruction (c). Ranges represent standard error of the mean.



(a)

(b)

(c)

Figure 20: Comparison of student average grade anomalies for men and women for each course of interest for classes during pre-remote (a), remote (b), and post-remote instruction (c). Ranges represent standard error of the mean.

12.0 Grades and grade anomalies before, during, and after remote COVID-19 instruction for bioscience and health-related majors: Overall trends and gender inequities

12.1 Introduction

In the wake of the COVID-19 pandemic, many education researchers have been focusing on assessing differences between online and in-person courses regarding student learning outcomes and classroom equity [142–145]. There are mixed findings regarding the effect of online instruction on student learning [142, 143, 260]. In this study, we explore overall trends in both grades and grade anomalies before, during, and after remote instruction due to COVID-19 in courses for bioscience and heath-related majors.

We define grade anomaly as the difference between a student's grade in a course of interest and their grade point average (GPA) in all other classes up to that point. The mean of this statistic for all students who took a course is the average grade anomaly (AGA). We divide average grade anomalies into "bonuses" and "penalties". A course in which students on average earn a lower grade than usual has an AGA with grade penalty, while a course in which students on average earn a higher grade than usual has an AGA with grade bonus.

Within our framework, we posit that grade anomaly may allow us to track, through institutional grade data, an important measure of how courses may affect students' academic self-concept. Academic self-concept is a relatively stable measure of a students' perceived ability to succeed in the academic sphere, and is based off of grades and outside feedback (e.g., from parents, peers, and instructors) [78–80, 233, 236]. Grades inform academic self-concept as both an external ("How good at math am I compared to other students?") and internal ("How good am I at math compared to English?") frame of reference [79, 80, 233]. We also note that, while academic self-concept is generally quite stable, it can change quite quickly during periods of transition (such as the transition from high school to university) [233]. Grade penalties in STEM courses during the first two semesters (but not later semesters) of university were negatively correlated with completing a STEM degree, even when controlling for gender, race, high school preparation, and college performance [239]. These findings hint at the importance of monitoring and minimizing grade penalties in students' first few semesters.

Our framework uses grade penalty as a central construct instead of grade because students' academic self-concept is often based on comparisons, not absolute grades [78]. Students may compare their grades across courses to determine which disciplines they excel at or struggle with [78]. Additionally, students tend to have a fairly fixed view of what "kind" of student they are, e.g., students may endorse the idea that "If I get As, I must be an A kind of person. If I get a C, I am a C kind of person" [53]. Grade anomalies may challenge or reinforce students' ideas about what kind of student they are, and if they are capable of succeeding in their chosen major. Many students who leave STEM majors explicitly cite lower grades than they are used to as a reason for doing so [53,54]. Grade penalties are more common and extreme in STEM disciplines than in humanities or social science departments [53,152,234,235].

Additionally, gender differences in performance and persistence in science, technology, engineering, and mathematics (STEM) have been closely studied in fields such as physics or engineering, in which women are underrepresented [?, 4, 7, 11, 12, 41, 45, 65, 115]. This line of research has been less common in fields such as biology, in which women are not underrepresented [120, 240]. However, even if fields in which women and men earn similar numbers of undergraduate degrees [1], women and men may have very different classroom experiences [?, 56, 188, 250].

There are many examples of gender inequities in biological science classrooms as well as career paths. Women in biology classrooms are less likely to participate in classroom discussions, are viewed as less knowledgeable by their peers, and tend to have lower exam grades in introductory biology courses [94, 182, 251]. After graduation, women with biological science graduate degrees are less likely than men to work as scientists after receiving graduate degrees [253]. If they pursue jobs in research, women in biology tend to have shorter publishing careers and lower yearly publishing rates than men in the same field [217,256]. If they choose to purse medical careers, women may still experience inequities: there are gender disparities in compensation and time to promotion for all academic medical specialties [254] as well as for physicians [255]. If gender differences in career outcomes are not explained by a lack of representation in the classroom, we hypothesize that academic self-concept may provide some insight [79, 80, 233]. Low academic self-concept may lead to lower future achievement and persistence because it discourages student engagement in a domain [79]. When women leave STEM disciplines, they often do so with higher grades than the men who remain in the program [39, 53, 54].

Broadly, in this research we aim to understand differences in students' grade anomalies before, during, and after remote instruction due to COVID-19 with a particular focus on gender differences in grades and grade anomalies. This will build on previous work which observed grade anomalies at this same institution for over ten years pre-COVID [?]. We aim to answer the following research questions regarding grade anomalies:

RQ1. Do grades or grade anomalies differ between before COVID, during remote

COVID teaching, and after remote COVID teaching?

RQ2. Are there gender differences in grades or grade anomalies, and do they differ between before COVID, during remote COVID teaching, and after remote COVID teaching?

12.2 Methodology

12.2.1 Participants

Participants in this study were enrolled in bioscience or health majors at a large, public, and urban institution. The student major breakdown can be found in Table **??**. All majors except for Neuroscience, Pharmacy, and Rehabilitation Science are offered through the Department of Biological Sciences. These students were chosen because of their similar course requirements, especially for large introductory science courses.

Grade data were collected over four years. We divide these semesters into three groups. First is "pre-remote" teaching, which consisted of the four semesters before instruction became remote due to the COVID-19 pandemic. Second is "remote" teaching, which spanned two semesters. Spring 2020 data were not included in any group because it included both in-person and remote instruction. Third was "post-remote" instruction, which covers two semesters. Additionally, we excluded courses that were taken during the summer semester. We excluded summer courses because they are not a typical representation of courses at our institution. For example, many summer students do not primarily attend our institution, but are local students visiting home for the summer. In addition the class sizes are an order of magnitude

Table 50: Major information for study participants. Undeclared was used as a potential major for students. This is because students are not required to declare a major until their third year. Undeclared students who later went on to declare any major not on this list were excluded.

Major	Pre-remote	Remote	Post-remote
Biological Sciences	34%	28%	22%
Bioinformatics/Computational Biology	2%	2%	1%
Ecology and Evolution	2%	1%	1%
Microbiology	6%	4%	3%
Molecular Biology	5%	5%	4%
Neuroscience	25%	19%	15%
Pharmacy	1%	0%	0%
Rehabilitation Science	6%	4%	3%
Undeclared	20%	36%	50%

smaller than those in the Fall and Spring semesters.

This left us with 14,152 pre-remote classes, 7,795 remote classes, and 6,143 post remote classes. We measure classes instead of students because most students take multiple courses studied over several years, so one student may be included multiple times in each data set. Demographic information for the student sample can be found in Table **??**. This research was carried out in accordance with the principles outlined in the University of Pittsburgh Institutional Review Board (IRB) ethical policy, and de-identified demographic data were provided through university records.

Table 51: Demographic information for study participants. Several survey options for ethnicity were excluded because they each made up less than 0.5% of the sample. These groups are Indigenous American, Pacific Islander, Not Specified, and Other. Unknown indicates that a student did not submit a response to the item, while Not Specified indicated that they chose the option “I prefer not to specify”.

	Sex		Race/Ethnicity					
Group	Female	Male	Asian	Black	Latine	Multiracial	White	Unknown
Pre-Remote	62%	38%	28%	5%	6%	5%	54%	0%
Remote	63%	37%	28%	5%	5%	5%	56%	1%
Post-Remote	62%	38%	25%	5%	5%	5%	59%	1%

12.2.2 Course Selection

We chose to study courses that were taken by the largest number of students, excluding non-major electives (for example, ”Introduction to Piano” or ”Public Speaking”) and courses that make up general education requirements. Thus, many courses were mandatory for students in the majors we focus on. However, not all courses were required for students in all the majors in our sample. Information about if a course was required, optional (i.e., an elective that count towards the major), or not required is included in Table 37 in Appendix H. The courses we chose are listed in Table 75 in Appendix H, along with information about the year in which the students typically take the course. We would like to note that, though it is not required for most majors studied, Human Physiology met our criteria because it is a commonly-chosen elective for both Biology and Rehabilitation Science Students.

Table 52: Grades and GPA points for this university's grading standards. For most majors, a C or above is a passing grade.

Grade	A/A+	A−	B+	B	B−	C+	C	C−	D+	D	D−	F
GPA Value	4.00	3.75	3.25	3.00	2.75	2.25	2.00	1.75	1.25	1.00	0.75	0.00

12.2.3 Measures

12.2.3.1 Course Grade

Course grades were based on the 0-4 scale used at our university, and a conversion of letter grades to GPA points can be seen in Table 52. We are unable to report grading schemes of each instructor, type of course (i.e., traditional lectures or active learning), or any other detailed course-level information due to the large number of courses sampled.

12.2.3.2 Grade Anomaly

GA was found by first finding each student's grade point average excluding the course of interest (GPA_{exc}). This was done by using the equation

$$GPA_{exc} = \frac{GPA_c \times Units_c - Grade \times Units}{Units_c - Units} \tag{14}$$

where GPA_c is the student's cumulative GPA, $Units_c$ is the cumulative number of units the student has taken, $Grade$ is the grade the student received in an individual course, and $Units$ is the number of units associated with an individual course.

After finding GPA_{exc} we can calculate grade anomaly (GA) by finding the difference between a student's GPA_{exc} and the grade received in that class:

$$GA = Grade - GPA_{exc}. \tag{15}$$

A negative GA corresponds to a course grade lower than a students' GPA in other classes and we call this a "grade penalty". A positive GA corresponds to a course grade higher than a students' GPA in other classes and we call this a "grade bonus". Average grade anomaly (AGA) is the mean of students' grade anomalies (GA) for each course, and is the metric by which we compare courses.

12.2.3.3 Analysis

To characterize both average grade anomaly (AGA) and grades, we found the sample size, mean, standard deviation, and standard error of each measurement for each course of interest. We calculated these statistics for women and men separately, and then for all students combined. We also compared the effect size of gender on both grade and grade anomaly, using Cohen's d to describe the size of the mean differences and unpaired t-tests to evaluate the statistical robustness of the differences. Cohen's d is calculated as follows:

$$d = \frac{\mu_1 - \mu_2}{\sqrt{\sigma_1^2 - \sigma_2^2}} \tag{16}$$

where μ_1 and μ_2 are the means of the two groups, σ_1 and σ_2 are the standard deviations [108] and Cohen's d is considered small if $d \sim 0.2$, medium if $d \sim 0.5$, and large if $d \sim 0.8$ [140]. We used a significance level of 0.05 in the t-tests as a balance between Type I (falsely rejecting a null hypothesis) and Type II (falsely accepting a null hypothesis) errors [108]. All analysis was conducted using R [174], using the package plotrix [245] for descriptive statistics, lsr [246] for effect sizes, and ggplot2 [247] to create plots.

12.3 Results

The primary result from our study in regards to research question 1 is the following: in general, course grades were highest during COVID-19 remote instruction compared to grades before or after, which can be seen in Figure 21. Additionally, grades after COVID-19 remote instruction tended to be the same or slightly lower after COVID-19 remote instruction. If average grades were lower after remote instruction, they tended to only be part of a letter grade lower (for example, if a pre-remote instruction average grade was a C+, it may be a C after remote instruction), which can be seen in either Figure 21 or in Table 76 in the Appendix I.

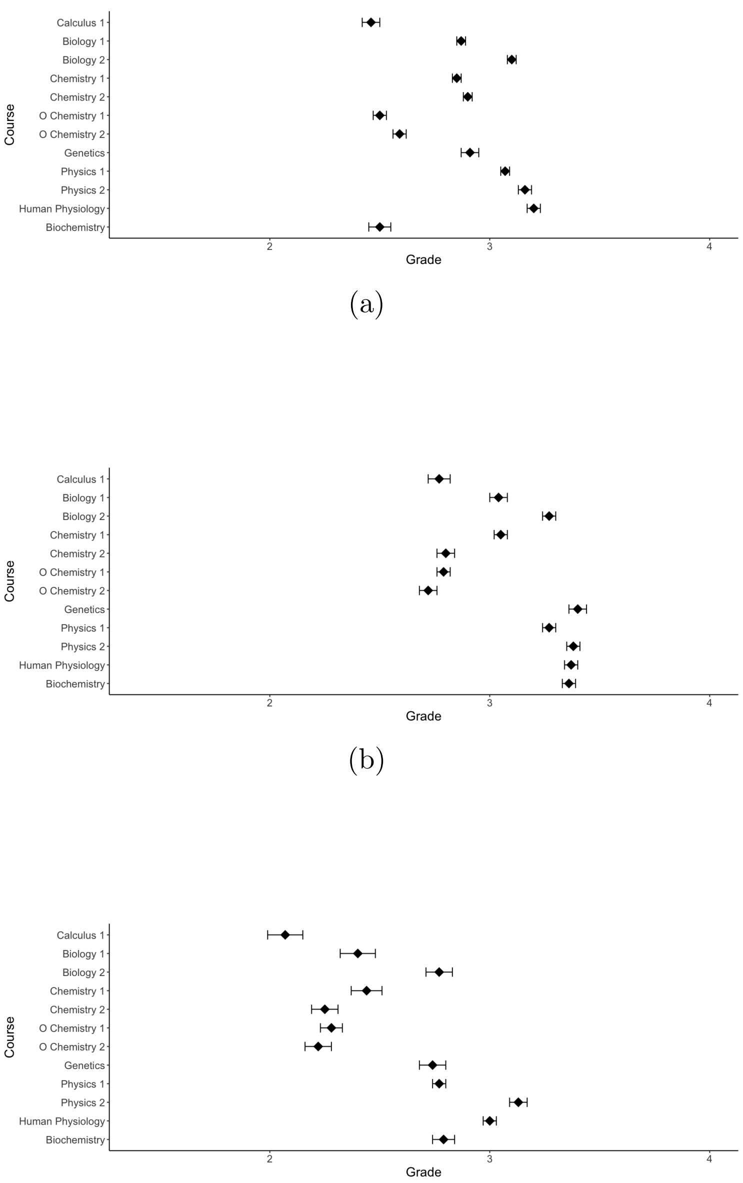
Also regarding research question 1, AGAs were the smallest during remote classes, which can be seen in Figure 22 or Table 77 in Appendix I. There were mixed trends regarding the relationships between pre and post remote AGAs. Most courses had larger AGAs post-remote instruction.

We do note that for any classes that are typically taken after a student's first semester (which are listed in Table 75), AGAs are expected to be higher. This is because grades tended to be higher during remote instruction than after. For example, student X and student Y both get a B− in their Biology 2 class. Student X took Biology 2 before the pandemic. Student X also took Biology 1 before the pandemic and got a B−, which was the average grade for pre-remote instruction. Thus, this B− has grade penalty compared to Biology 1. Student Y took Biology 2 in-person during post-remote instruction. Student Y took Biology 1 during remote instruction and got a B+, which was the average grade during remote instruction. To student Y, the B− in Biology 2 is a grade penalty compared to Biology 1. This example can be generalized to explain why AGAs tended to be larger during post remote instruction versus pre-remote instruction.

There were generally more grade differences in AGAs than in grades. For example, before remote instruction, there were gendered grade differences in two courses (Calculus 1 and Genetics, which can be seen in Figure 23a), in which women tended to have higher grades than men. During COVID-19 remote instruction, there was only one gendered grade difference (in Physics 2, favoring men). During post-remote instruction, all gendered grade differences (in Biology 1, as well as Organic Chemistry 1 and 2, which can be seen in Figure 23c) favored men. However, most courses did not have gendered grade differences. This trend is very different than gendered AGA differences. Before the pandemic, men had smaller grade penalties than women in six of the twelve of the courses studied, while women had smaller grade penalties in only one course, Genetics. During remote instruction, men had smaller grade penalties than women in three courses, and after remote instruction, men had smaller grade penalties than women in five courses. There were no courses in which women had smaller grade penalties than men during or after remote instruction. In most classes, men and women had indistinguishable AGAs. However, in most first-year courses, women had larger average grade penalties than men. Below, we investigate trends in specific subjects regarding student grades and AGAs.

12.3.1 Biology Courses

Biology grades were typically in the upper half of the course grade distributions, before, during, and after remote instruction. This means that, among the courses

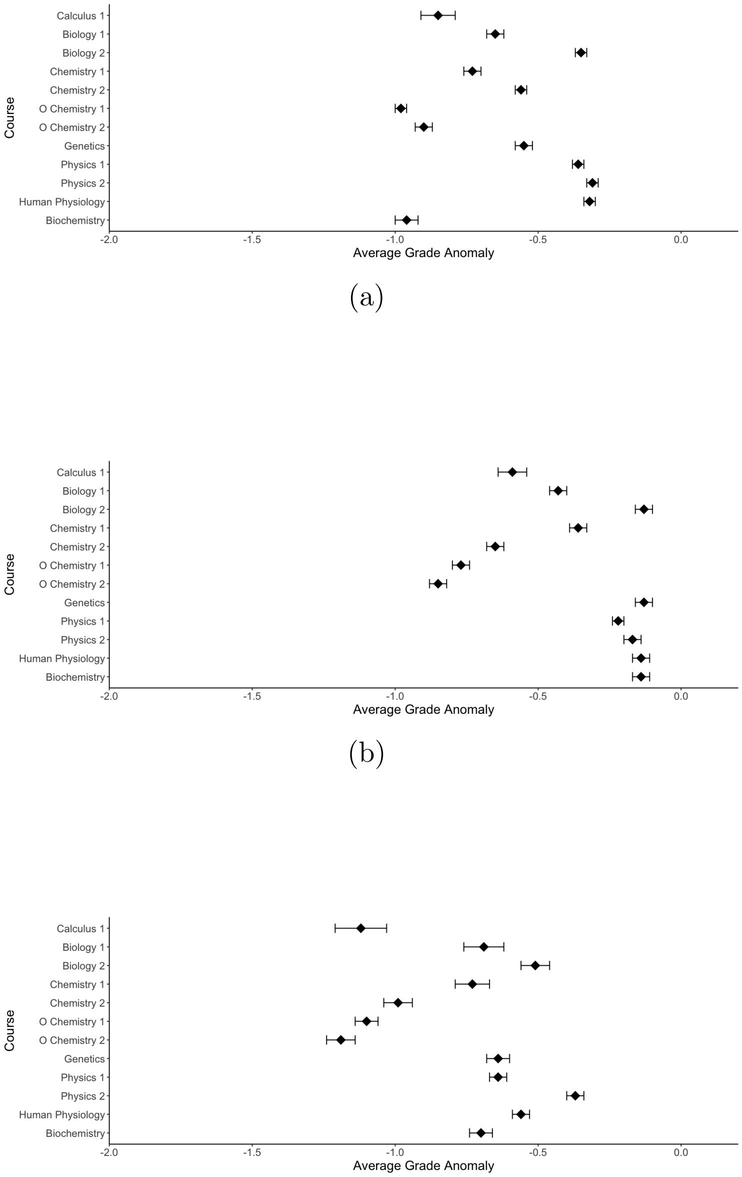


(a)

(b)

(c)

Figure 21: Average grades for each course of interest for classes before (a), during (b), and after (c) remote instruction. Ranges represent standard error of the mean.

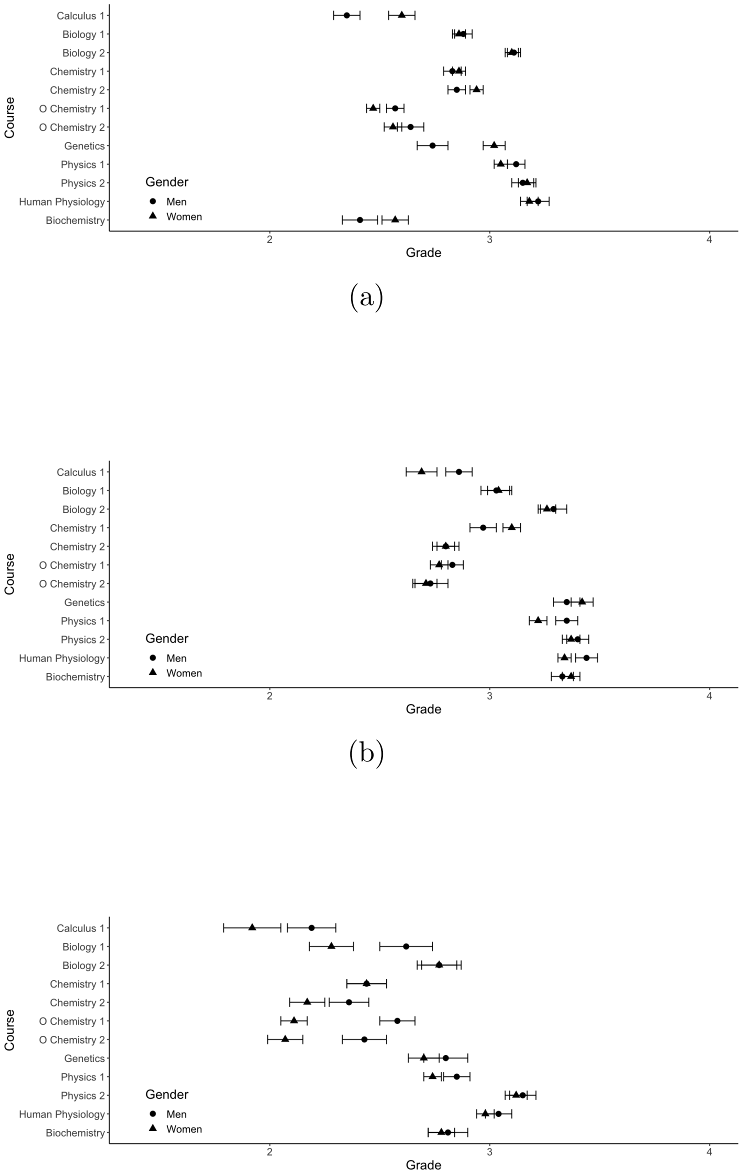


(a)

(b)

(c)

Figure 22: Average grade anomalies for each course before (a), during (b), and after (c) remote instruction. Ranges represent standard error of the mean.

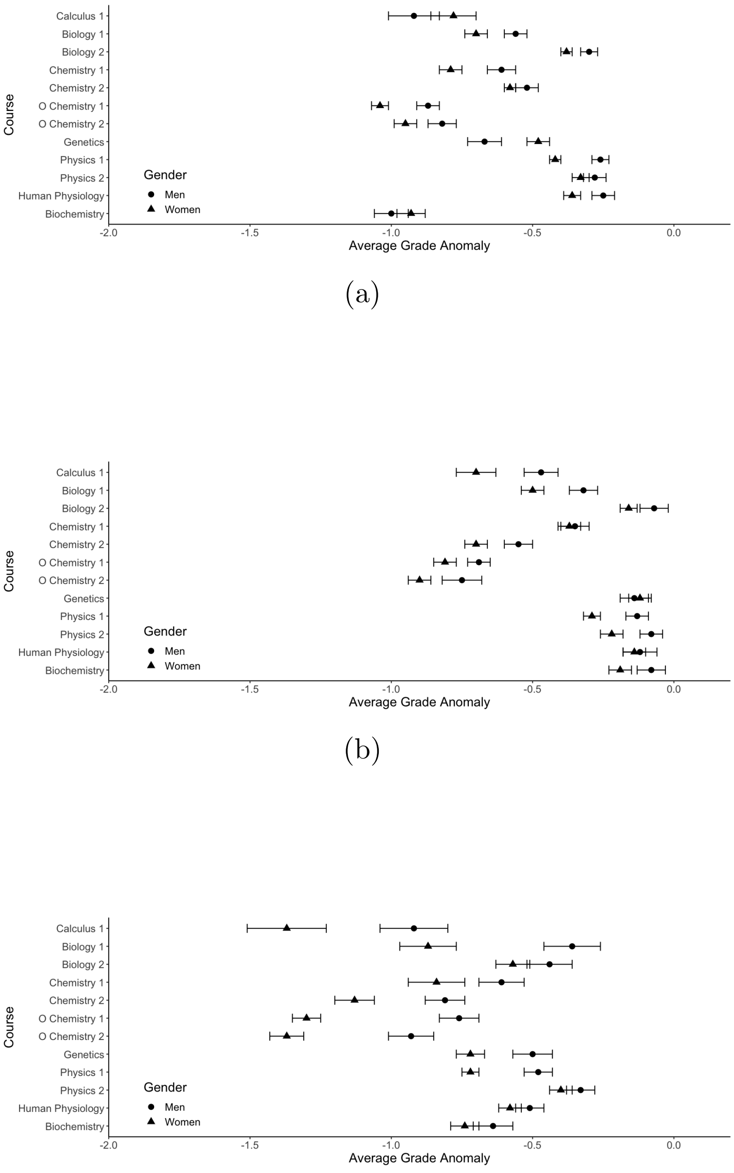


(a)

(b)

(c)

Figure 23: Comparison of average grades for men and women for each course of interest before (a), during (b), and after (c) remote instruction. Ranges represent standard error of the mean.



(a)

(b)

(c)

Figure 24: Comparison of average grade anomalies for men and women for each course of interest before (a), during (b), and after (c) remote instruction. Ranges represent standard error of the mean.

included in this study, students with bioscience and health-related majors tended to have their highest grades in courses offered by the Biological Science (Biology) department, which can be seen in Figure 21. Generally, Biology grades during remote instruction were higher than before or after. Course grades after remote instruction tended to be the same or one partial letter grade lower than they were before remote instruction (for example, in Biology 1 had an average course grade of B− before remote instruction, but C+ after).

However, some Biology classes had gender differences in grades, which can be seen in Table 53. Genetics had the largest gender difference in course grades ($d = 0.26$, $p < 0.01$) among the courses studied pre-remote instruction (favoring women), and Biology 1 had a statistically significant gender difference in course grades ($d = 0.30$, $p < 0.05$) post-remote instruction (favoring men).

AGAs for Biology courses also tended to be average or small compared to other courses studied, as shown in Figure 22. Though Bioscience students can generally expect lower grades in biology compared to most courses they take, the grade penalties are relatively modest compared to other STEM courses included in this study. Biology course AGAs tended to have the smallest magnitude during remote instruction, as seen in Table 53. AGAs tended to be either similar or slightly larger after post-remote instruction than before remote instruction. One exception to this trend was Biochemistry, which had a smaller grade penalty after than before remote instruction.

Most Biology courses have a gender difference in AGAs during at least one time in the study (see Table 53). All of these gender differences showed that women had larger grade penalties on average than men. Biology 1, Biology 2, Genetics, and Human Physiology had small ($d \sim 0.2$) gender differences pre-remote teaching, and Biology 1 had a small gender difference during remote instruction. Biology 2 and Genetics did not have grade penalty gender differences during or after remote instruction. Genetics did not have any gender differences during remote instruction, and returned to a small gender difference after. Biology 1 grade anomaly gender differences became larger over time, and during post-remote instruction there was a medium grade difference ($d \sim 0.5$) between men and women.

12.3.2 Chemistry Courses

Chemistry courses tended to have some of the lowest grades among any courses studied, shown in Figure 21. Notably, Organic Chemistry 2 had the lowest average grade of the courses studied before and during remote instruction, and had the second lowest average grade after. Figure 21 also shows that Organic Chemistry 1 had similar grades, though they were slightly higher than those in Organic Chemistry 2. Introductory Chemistry 1 and 2 had grades that were closer to the average of courses studied before and after remote instruction, but Chemistry 2 had very low average grades, comparable to Organic Chemistry 1. This means that, among the courses included in this study, students with bioscience and health-related majors tended to have their lowest grades in courses offered by the Chemistry department. Generally, grades during remote instruction were higher than before or after, as seen in Figure 21. Course grades after remote instruction tended to be the same or one partial letter grade lower than they were before remote instruction. Before and during remote instruction, Table 54 shows that no Chemistry courses had gender differences in grades. However, during post-remote instruction, men tended to have higher grades in both Organic Chemistry 1 and 2.

AGAs for Chemistry courses also tended to be large compared to other courses studied, as shown in Figure 22. Chemistry course grade anomalies tended to have the smallest magnitude during remote instruction. AGAs tended to be either similar or slightly larger during post-remote instruction than before remote instruction, which can be seen in Table 54. Notably, after remote instruction, students taking Organic Chemistry 1 or 2 can expect to receive a grade over one full letter grade lower than their GPA excluding these classes.

Men had smaller grade anomalies in all Chemistry classes except Chemistry 2 before remote instruction (see Table 54), and the gender differences tended to be small ($d \sim 0.2$). Chemistry had a small gender difference in grade anomalies during remote instruction, but no statistically significant difference before or after. Men tended to have smaller grade penalties than women with a medium effect size ($d \sim 0.5$) in both Organic Chemistry 1 and 2.

12.3.3 Math Courses

Before remote instruction, Table 55 shows that Calculus 1 had an average grade of C+. During remote classes, Calculus 1 had an average grade of B−. During Post-Remote classes, the average grade was C, which is the lowest grade of any course studied, which can be seen in Figure 21. Calculus 1 had a statistically significant grade difference pre-remote instruction, in which women tended to have higher grades than men. Calculus 1 AGAs tended to be the largest after Organic Chemistry 1 and 2 (as well as Chemistry 2 during remote instruction), which can be seen in Figure 22. Calculus 1 AGAs were -0.85 before, -0.59 during, and -1.12 during post-remote instruction. This means that after remote instruction, students could expect to have a Calculus grade more than one full letter grade lower than their GPA in other courses. There were no gender differences in AGAs before the pandemic, but women had larger AGAs than men during and after remote instruction.

12.3.4 Physics Courses

Before, during, and after remote instruction, Physics course grades were generally higher than those in other subjects, which can be seen in Figure 21. For both Physics 1 and 2, Table 56 shows that grades were highest during remote instruction. For Physics 1, the average grade before remote instruction was also a B, but dropped to a B− after remote instruction. For Physics 2, the average grade both before and after remote instruction was a B. Both Table 56 and Figure 23 show that neither Physics 1 or 2 had gendered grade differences before or after remote instruction. However, during remote instruction, men tended to have higher grades than women, with a small effect size ($d \sim 0.2$).

AGAs in Physics courses also tended to have a smaller magnitude than those of other subjects (see Figure 22). Grade anomalies for both Physics courses were smallest during remote instruction, but in most cases, student's physics grades were less than half a letter grade lower than their GPA excluding those courses, which can be seen in Table 56. The exception to this trend is that during post remote instruction, Physics 1students tended to have a grade penalty between half and three-fourths of a letter grade. Physics 1 courses had small ($d \sim 0.2$) AGA gender differences favoring men before, during, and after remote instruction. Physics 2 had no gender difference in AGA before or after remote instruction, but men tended to have smaller grade penalties than women during remote instruction, with a small effect size ($d \sim 0.2$).

12.4 Discussion

12.4.1 Do grades and AGAs differ between before COVID, during remote COVID teaching, and after remote COVID teaching?

Grades are important to students for a variety of reasons such as continuing their major, scholarship requirements, graduate school admissions. Broadly, grades were higher during the transition to remote instruction and were lower again during post-remote instruction. AGAs, unlike grades, do not have a direct effect on students' outcomes such as scholarships and graduate admissions. A student with an A average who receives a B in a class has the same grade anomaly with a B average who receives a C in the class. Here, we use the idea of academic self concept form Situated Expectancy Value Theory to frame how students may think about grade anomalies. Students tend to have a somewhat fixed idea of what sort of student they are (for example, they may endorse that idea that "If I get As, I must be an A kind of person") [53]. AGAs may challenge a student's idea about what kind of student they are. In particular, students may compare their grades across courses to determine which disciplines they excel at or struggle with [78].

During the transition from pre-remote to remote instruction, there tended to be large increases in average grades (often over one full letter grade) for classes students usually take in their third year of university or later (see Table 75 in Appendix K for information about when students tend to take each course). On the other hand, Chemistry 1 and Biology 1 courses, which are primarily taken by first-semester students, had worse grades during remote instruction than during pre-remote instruction. These increases in grades may be due to a range of factors. For example, grading schemes and assessment types may have been changed, or instructors may have been more flexible for students than during pre-remote classes [258]. We note that student performance on content-based surveys is similar in online and in-person administration [146], and answer-copying on homework does not significantly differ between remote and in-person instruction [147]. This leaves open the possibility that grade differences between in-person and online courses are not inherent, but may be the result of instructor choices in class policies.

After the transition to remote instruction, AGAs for almost all courses were lower. That is, students' grades were more consistent during remote instruction, so that most classes deviated less from a students average GPA during remote instruction. We hypothesize that these smaller grade anomalies may result in students being less concerned that they can succeed in their discipline, and may rely more on other factors (such as interest) to make decisions regarding major and career choice.

Next, we discuss changes from remote to post-remote courses. Comparing remote to post-remote classes, only two courses had better grades during in-person courses were Biology 1 and Chemistry 1, which are classes primarily taken by first-year students. During the transition back to in-person classes, all courses developed larger AGAs. Part of this drop is due to the fact that grades were generally higher during remote than in-person classes. However, this trend also holds for many first-year classes (in which students would have no grades from remote instruction), so there are likely other factors involved.

Trends in grades between pre- and post-remote instruction were more complicated. Half of the courses studied had the same or a higher average grade during post-remote courses, and half had a lower grade during post-remote than pre-remote courses. Again, first-semester courses (Calculus 1, Chemistry 1, and Biology 1) had worse outcomes during pre-remote than post-remote instruction, and generally courses that students took in their first two years were more likely to show grade decreases from pre- to post-remote instruction.

There was no course in which AGAs decreased in magnitude from pre to post, and almost all courses had larger AGAs during post-remote than pre-remote courses. However, several courses had similar AGAs between pre and post, including two first-year courses: Biology 1 and Chemistry 1. While it is encouraging that the AGAs are not getting worse over time for these courses, it should still be noted that the AGAs are on average larger than they are for courses students take later in their major.

The courses with the lowest grades and largest AGAs over all time periods are Calculus 1 and Organic Chemistry 1 and 2, which are often labeled as "weed out" courses. Grade penalties are more common and larger in STEM diciplines than in social sciences or humanities [53, 152, 234, 235], but our finsdings show that there are significant variations in AGAs even among STEM courses. In general, Biology and Physics courses tended to have had the smallest AGAs, while Math and Chemistry courses tended to have the largest. Thus, AGAs are not a simple issue of STEM courses having larger AGAs than non-STEM courses. Instructors and departments with comparatively lower grades and larger AGAs than others may benefit from pedagogies implemented by other STEM departments and instructors at their institution.

There are likely to be a range of potential factors contributing to differences in grade and AGAs over time. Though there is a possibility that some students are cheating, cheating on exams seems to have only small increases in the USA during the pandemic, though the effect may be larger in other regions [259]. There is also research that suggests that there are specific factors that could lead to increased grades during remote instruction. For example, because there were were more low-stakes assessments during remote instruction, students may be more likely to engage in spaced practice instead of "cramming" for assessments during remote learning [260]. One study showed that students had higher grades during COVID-19 remote instruction even on identical assessments that were also given online pre-pandemic [260]. One study that focused on quantum mechanics (an upper-level physics course) found that implementing low-stakes formative assessments instead of exams did not lead to lower scores on course post-test (which only contributed to a small amount of the students' final grade) [143]. Though these studies do not specify any specific reason that there may be differences in grades during remote versus in-person classes, they do suggest that the increase in grades do not necessarily correlate with lowered academic standards or cheating.

One of the concerning trends from this study was that there are lower grades and larger AGAs for students in their first two years of university than for later years. Low grades early on during the transition to university are particularly concerning. Low grades, even if they are high enough to continue in a major, are a common reasons that students cite for leaving a STEM major [53, 54]. Additionally, because academic self-concept is most in flux during transitional periods, low grades early in a student's college career may negatively affect students' self-concept more than low grades received in later years [233].

We also note that students tended to have lower grades and larger grade anomalies during post-remote courses than pre-remote courses. This was more true for courses that students tended to take after their first two years of university. We hypothesize that this is because students who took introductory level courses did not have the same supports (such as first year programs that focus on building community) as students who started university during in-person instruction. It is possible that students who started university during remote courses may not have developed a sense of community and group study skills that students who started during in-person classes may have. This may result in less preparation for courses students take later in their major which are often a higher level and more complex.

12.4.2 Are there gender differences in grades or grade anomalies, and do they differ between before COVID, during remote COVID teaching, and after remote COVID teaching?

Before remote instruction, there were either no statistically significant grade differences, or women had higher grades than men. However, during and after remote instruction, all classes with statistically significant grade differences favored men. In each case, there were only a few courses with any statistically significant grade differences. One particularly concerning class was Calculus 1. Before and during remote instruction, women had an average grade of C+, but an average grade of C− during post-remote instruction. This means that on average, women did not have a grade in Calculus 1 needed to continue in the major. These women need to choose between taking the class again, which involves a commitment of both time and tuition, or change to a major that does not require Calculus. Though other courses had a larger gender differences, this is the only course and group for which the average outcome was not passing the course.

There were many courses that had statistically significant gendered AGA differences. In all of these courses (except pre-remote Genetics), men had smaller AGAs than women. The largest of these AGA differences were post-remote Biology 1, Organic Chemistry 1, and Organic Chemistry 2. For women in bioscience majors, a large grade anomaly in thier first biology course at university may be particularly concerning, and potentially lead them to believe that they do not "have what it takes" to succeed in her major. Women often report worrying more then men that they do not understand the material even if they receive A’s, B’s, or C’s (which are grades that allow students to continue in most programs) [53, 115]. This trend has been found to be particularly strong among high-achieving women [53].

We hypothesize that women may be more likely to have a low academic-self-concept than men at similar performance levels. Prior work has theorized that men are more likely to separate their grades and sense of academic self-concept [53,54,237]. Academic self-concept is formed through grades and feedback from outsiders. Women are generally less likely to receive recognition from instructors [56,130,208], so women may rely more than men on grade information to develop their academic self-concept [53, 54, 237]. Women also tend to earn higher grades than men who have the same standardized test scores [53, 203], so they may be more accustomed to higher grades. As a result, they may have more concern about grades that are lower than what they are accustomed to, or they may compare their relatively-low STEM grades and view themselves as less able to succeed in biology sciences or a health-related field than a subject that gives them the recognition for their work that they are accustomed to [53, 54].

AGAs and raw grade data do not always reveal the same trends: there are many more gender differences in AGA than in grades in the findings presented here. This trend reveals how AGA may be a useful measure. For example, an instructor may not see any gender differences in grades, which is one important indicator of gender equity. However, if they do not know the gender differences in AGA, an instructor or department may not recognize how those grades may be perceived by women and men in their classes. Understanding both grades and AGA differences may allow instructors to understand classroom-level inequities better.

12.5 Conclusion and Future Research

In this work we found that grade penalties exist for all the courses we studied. Over the transition from in-person to remote courses, grades were higher and grade penalties were smaller. During the transition back to in-person instruction the grades were lower and grade penalties became larger again. Further, there were more gender differences in both grades and grade penalties (favoring men) after remote teaching than before or during.

These results are very important because they provide evidence that courses in STEM departments tend to have grade penalties, and that these penalties tend to decrease during remote instruction. Additionally, AGA may also act as a useful measure of academic self-concept that is easy for institutions to access.

Although we have evidence of grade penalties in the studied courses as well as gendered grade anomaly differences, we did not have access to syllabi or other information about individual courses offered over the period of data collection. Therefore, we are not able to pinpoint specific practices that may lead to grade penalties, grade bonuses, or gender inequities at our institution.

Finally, this research is based at a primarily white, large, public university. While our results may generalize to similar institutions, we do not know what patterns of grade anomalies exist at smaller liberal arts colleges, minority-serving institutions, or community colleges in the US. Additionally, it may also be useful to repeat similar research in other countries, as many countries worldwide were affected differently by the COVID-19 pandemic.

Table 53: Means and standard deviations (SD) of grades and grade anomalies by gender for courses offered by the Biological Science Department before (Pre), during (Rem), and after (Post) remote instruction. Human Physiology (Physiology) and Biochemistry (Biochem) are abbreviated. Cohen's d is positive if men had higher grades or smaller AGAs than women in a course. A bold Cohen's d signifies that a t-test showed significant differences between men and women. $^{\gamma} = p < 0.05$, $^{\beta} = p < 0.01$, and $^{\alpha} = p < 0.001$.

		Women					Men						
			AGA		Grade			AGA		Grade		Cohen's d	
Course	Type	N	Mean	SD	Mean	SD	N	Mean	SD	Mean	SD	AGA	Grade
Biology 1	Pre	1079	-0.70	1.17	2.86	1.02	569	-0.56	0.92	2.88	1.00	**0.13**$^{\gamma}$	0.02
	Rem	326	-0.50	0.74	3.04	0.83	185	-0.32	0.70	3.03	0.91	**0.25**$^{\gamma}$	-0.01
	Post	141	-0.87	1.15	2.28	1.22	76	-0.36	0.87	2.62	1.03	**0.48**$^{\alpha}$	**0.30**$^{\gamma}$
Biology 2	Pre	902	-0.38	0.65	3.10	0.77	519	-0.30	0.64	3.11	0.79	**0.12**$^{\gamma}$	0.02
	Rem	363	-0.16	0.65	3.26	0.80	186	-0.07	0.62	3.29	0.79	0.14	0.03
	Post	121	-0.57	0.65	2.77	0.85	91	-0.44	0.72	2.77	0.98	0.20	0.00
Genetics	Pre	453	-0.48	0.78	3.02	0.96	305	-0.67	1.01	2.74	1.18	**-0.22**$^{\gamma}$	**-0.26**$^{\beta}$
	Rem	267	-0.12	0.62	3.42	0.80	172	-0.14	0.63	3.35	0.84	0.04	0.08
	Post	225	-0.72	0.80	2.7	1.02	138	-0.50	0.79	2.8	1.14	**0.27**$^{\gamma}$	0.10
Physiology	Pre	571	-0.36	0.76	3.18	0.90	334	-0.25	0.69	3.22	0.87	**0.14**$^{\gamma}$	0.05
	Rem	480	-0.14	0.61	3.34	0.75	230	-0.12	0.67	3.44	0.78	0.03	0.13
	Post	527	-0.58	0.80	2.98	0.96	282	-0.51	0.82	3.04	1.00	0.10	0.06
Biochem	Pre	305	-0.93	0.88	2.57	1.03	219	-1.00	0.92	2.41	1.15	-0.09	-0.15
	Rem	349	-0.19	0.67	3.37	0.81	234	-0.08	0.68	3.33	0.82	0.16	-0.05
	Post	376	-0.74	1.02	2.78	1.24	199	-0.64	0.98	2.81	1.26	0.10	0.03

Table 54: Means and standard deviations (SD) of grades and grade anomalies by gender for courses offered by the Chemistry Department before (Pre), during (Rem), and after (Post) remote instruction. Chemistry (Chem) and Organic Chemistry (Orgo) are abbreviated. Cohen's d is positive if men had higher grades or AGAs than women in a course. A bold Cohen's d signifies that a t-test showed significant differences between men and women. $^{\gamma} = p < 0.05$, $^{\beta} = p < 0.01$, and $^{\alpha} = p < 0.001$.

		Women					Men						
			AGA		Grade			AGA		Grade		Cohen's d	
Course	Type	N	Mean	SD	Mean	SD	N	Mean	SD	Mean	SD	AGA	Grade
Chem 1	Pre	1160	-0.79	1.38	2.86	0.87	679	-0.61	1.41	2.83	0.97	**0.13**$^{\gamma}$	-0.04
	Rem	374	-0.37	0.70	3.10	0.74	224	-0.35	0.69	2.97	0.83	0.04	-0.17
	Post	134	-0.84	1.13	2.44	1.09	127	-0.61	0.88	2.44	1.02	0.23	0.00
Chem 2	Pre	970	-0.58	0.77	2.94	0.86	597	-0.52	0.86	2.85	0.91	0.07	-0.09
	Rem	359	-0.70	0.69	2.80	0.83	217	-0.55	0.65	2.80	0.90	**0.23**$^{\gamma}$	0.00
	Post	174	-1.13	0.91	2.17	1.08	138	-0.81	0.82	2.36	1.09	0.37	0.18
Orgo 1	Pre	1051	-1.04	0.99	2.47	1.12	583	-0.87	0.94	2.57	1.07	**0.18**$^{\alpha}$	0.09
	Rem	530	-0.81	0.79	2.77	0.95	270	-0.69	0.70	2.83	0.88	0.16	0.06
	Post	340	-1.30	0.91	2.11	1.09	197	-0.76	0.97	2.58	1.16	**0.58**$^{\alpha}$	**0.42**$^{\alpha}$
Orgo 2	Pre	682	-0.95	0.92	2.56	1.08	420	-0.82	0.98	2.64	1.17	**0.13**$^{\gamma}$	0.07
	Rem	441	-0.90	0.81	2.71	0.97	213	-0.75	0.90	2.73	1.12	0.18	0.02
	Post	249	-1.37	1.02	2.07	1.22	172	-0.93	1.08	2.43	1.28	**0.43**$^{\alpha}$	**0.30**$^{\beta}$

Table 55: Means and standard deviations (SD) of grades and grade anomalies by gender for courses offered by the Mathematics Department before (Pre), during (Rem), and after (Post) remote instruction. Cohen's d is positive if men had higher grades or AGAs than women in a course. A bold Cohen's d signifies that a t-test showed significant differences between men and women. $^{\gamma} = p < 0.05$, $^{\beta} = p < 0.01$, and $^{\alpha} = p < 0.001$.

		Women					Men						
			AGA		Grade			AGA		Grade		Cohen's d	
Course	Type	N	Mean	SD	Mean	SD	N	Mean	SD	Mean	SD	AGA	Grade
Calculus 1	Pre	379	-0.78	1.49	2.60	1.11	433	-0.92	1.91	2.35	1.3	-0.08	**-0.21**$^{\beta}$
	Rem	226	-0.70	0.95	2.69	0.98	222	-0.47	0.85	2.86	0.93	**0.26**$^{\gamma}$	0.17
	Post	131	-1.37	1.57	1.92	1.45	162	-0.92	1.55	2.19	1.44	**0.29**$^{\gamma}$	0.19

Table 56: Means and standard deviations (SD) of grades and grade anomalies by gender for courses offered by the Physics Department before (Pre), during (Rem), and after (Post) remote instruction. Cohen's d is positive if men had higher grades or AGAs than women in a course. A bold Cohen's d signifies that a t-test showed significant differences between men and women. $^{\gamma} = p < 0.05$, $^{\beta} = p < 0.01$, and $^{\alpha} = p < 0.001$.

		Women					Men						
			AGA		Grade			AGA		Grade		Cohen's d	
Course	Type	N	Mean	SD	Mean	SD	N	Mean	SD	Mean	SD	AGA	Grade
Physics 1	Pre	720	-0.42	0.61	3.05	0.77	438	-0.26	0.64	3.12	0.84	**0.26**$^{\alpha}$	0.09
	Rem	414	-0.29	0.60	3.22	0.78	295	-0.13	0.58	3.35	0.81	**0.26**$^{\beta}$	**0.16**$^{\gamma}$
	Post	590	-0.72	0.79	2.74	0.95	291	-0.48	0.76	2.85	0.98	**0.31**$^{\alpha}$	0.12
Physics 2	Pre	449	-0.33	0.61	3.17	0.78	292	-0.28	0.66	3.15	0.88	0.08	-0.02
	Rem	394	-0.22	0.63	3.37	0.71	253	-0.08	0.51	3.40	0.79	**0.23**$^{\gamma}$	0.05
	Post	384	-0.40	0.76	3.12	0.93	252	-0.33	0.73	3.15	0.98	0.10	0.03

13.0 Future Directions

The research presented here gives insight into the relationship between motivational beliefs and academic performance in physics courses. It also investigates gender differences in these constructs in a range of contexts, including introductory physics for engineers and physical science majors, introductory physics for bioscience majors, and upper-division physics courses. Further, the research presented here explores how remote teaching due the COVID-19 pandemic impacts student grades and grade penalties.

The present studies only focus on gender differences, and do not include any investigation focusing on other underrepresented groups. In future studies, it would be useful to carry out similar investigations considering more aspects of students' identities, such as race/ethnicity, disability status, first-generations status, and sexual orientation as well as how the intersectionality impacts students with multiple marginalized identities. Additionally, studies with more nuanced measures of gender can be useful to capture a diverse range of student experiences.

This research was carried out at a large public research university in the northeastern United States. Thus, the results may not always be realizable to other types of institutions, such as two-year colleges, minority serving institutions, or universities outside of the United States.

Future studies should also investigate students' motivational beliefs and academic performance in the classes in which there is an intentional focus on equity and inclusion or those using research-based classroom interventions with control groups to further study their effect on the perception of the inclusiveness of the learning environment and on students' course outcomes.

Appendix A. Factor Analysis and Predicting Low-Stakes Assessment Scores in introductory physics courses for engineering and physical science majors.

Table 57: Survey fit indices for confirmatory factor analysis for both Physics 1 and 2. Chronbach's α for pre and post Test Anxiety and Self-Efficacy is included. Students were included if they competed the pre or post survey.

	Fit Indicies				Cronbach's α			
					Self-Efficacy		Test Anxiety	
Course	CFI	TLI	RMSEA	SRMR	Pre	Post	Pre	Post
Physics 1	0.93	0.91	0.07	0.05	0.74	0.81	0.91	0.91
Physics 2	0.93	0.92	0.08	0.05	0.83	0.83	0.92	0.92

Table 58: Physics 1 and 2 low-stakes assessment scores predicted by student sex, High School GPA, SAT/ACT Math scores, pre or average self-efficacy and pre or average test anxiety. Standardized regression (β) coefficients are provided.
$^{*} = p < 0.05$, $^{**} = p < 0.01$, and $^{***} = p < 0.001$.

	Physics 1		Physics 2			
	Pre	Avg.		Pre		Avg.
Variable	Model 1	Model 1	Model 1	Model 2	Model 3	Model 1
Sex (M=0, F=1)	0.05	-0.04	0.08	0.08	0.10	0.02
High School GPA	0.22^{***}	0.31^{***}	0.23^{***}	0.25^{***}	0.24^{***}	0.12^{*}
SAT/ACT Math	0.05	0.05	0.03	0.04	0.03	0.05
Self-Efficacy	0.09	-0.05	0.04			-0.01
Test Anxiety	-0.10	-0.07	-0.10^{*}	-0.07		0.06
N	401	186	552	552	552	304
Adjusted R^2	0.06	0.07	0.07	0.08	0.08	0.01

Appendix B. Factor Analysis and Predicting Low-Stakes Assessment Scores in introductory physics courses for bioscience majors

Table 59: Survey fit indices for confirmatory factor analysis for both Physics 1 and Physics 2. Chronbach's α for pre and post Test Anxiety (TA) and Self-Efficacy (SE) is included. For Physics 1, $N = 516$ and for Physics 2, $N = 608$.

	Fit Indicies				Cronbach's α			
					Self-Efficacy		Test Anxiety	
Course	CFI	TLI	RMSEA	SRMR	Pre	Post	Pre	Post
Physics 1	0.93	0.91	0.07	0.05	0.74	0.81	0.91	0.91
Physics 2	0.93	0.92	0.08	0.05	0.83	0.83	0.92	0.92

Table 60: Physics 1 and 2 low-stakes assessment scores predicted by student sex, High School GPA (HS GPA), SAT/ACT Math scores, pre or average self-efficacy and pre or average test anxiety. Standardized regression (β) coefficients are provided. $^{*} = p < 0.05$, $^{**} = p < 0.01$, and $^{***} = p < 0.001$.

	Physics 1		Physics 2			
	Pre	Avg.		Pre		Avg.
Variable	Model 1	Model 1	Model 1	Model 2	Model 3	Model 1
Sex (M=0, F=1)	0.05	-0.04	0.08	0.08	0.10	0.02
HS GPA	0.22***	0.31***	0.23***	0.25***	0.24***	0.12*
SAT/ACT Math	0.05	0.05	0.03	0.04	0.03	0.05
Self-Efficacy	0.09	-0.05	0.04			-0.01
Test Anxiety	-0.10	-0.07	-0.10*	-0.07		0.06
N	401	186	552	552	552	304
Adjusted R^2	0.06	0.07	0.07	0.08	0.08	0.01

Appendix C. Survey Validation for Self-Efficacy and Test Anxiety Constructs Before, During, and After Remote Instruction

Acceptable cutoff values for the Comparative Fit Index (CFI) and Tucker Lewis Index (TLI) are ≥ 0.95 [106], and the acceptable cutoff the Root Mean Square Error of Approximation (RMSEA) and Standardized Root Mean Square Residual (SRMR) are both ≤ 0.08 [107]. The fit indices for our survey meet these standards and can be found in Table 61. Acceptable values for Cronbach's α in education are between 0.7 and 0.9 [108]. All constructs fell within this range, as seen in Table 61. Standardized factor loadings were all above 0.5 [106], which can be see in Table 62.

Table 61: Survey fit indices for confirmatory factor analysis for both Physics 1 and Physics 2. Additionally, Chronbach's α for pre and post Test Anxiety (TA) and Self-Efficacy (SE) is included.

	Fit Indicies				Cronbach's α			
Course	CFI	TLI	RMSEA	SRMR	Pre SE	Post SE	Pre TA	Post TA
Physics 1	0.92	0.90	0.07	0.06	0.70	0.77	0.89	0.83
Physics 2	0.93	0.92	0.07	0.04	0.82	0.82	0.89	0.90

Table 62: Survey items with standardized factor loadings. For Physics 1, $N = 786$ and for Physics 2, $N = 546$.

		Factor Loading			
		Physics 1		Physice 2	
	Construct Name/Item Text	Pre	Post	Pre	Post
	Self-Efficacy				
1.	I am able to help my classmates with physics in the laboratory or in recitation	0.59	0.68	0.69	0.72
2.	I understand concepts I have studied in physics	0.66	0.70	0.75	0.77
3.	If I study, I will do well on a physics test	0.64	0.66	0.82	0.77
4.	If I encounter a setback in a physics exam, I can overcome it	0.56	0.68	0.73	0.77
	Test Anxiety				
5.	I am so nervous during a physics test that I cannot remember what I have learned	0.82	0.80	0.79	0.79
6.	I have an uneasy, upset feeling when I take a physics test	0.87	0.67	0.90	0.92
7.	I worry a great deal about physics tests	0.81	0.62	0.82	0.77
8.	When I take a physics test, I think about how poorly I am doing	0.80	0.86	0.79	0.85

Appendix D. Regression models predicting high-stakes assessments Before, During, and After Remote Instruction

Table 63: Physics 1 high-stakes assessment scores predicted by student sex, HS GPA, SAT/ACT Math scores, pre self-efficacy and pre test anxiety. For virtual classes, $n = 216$ and for in-person classes, $n = 426$. Standardized regression coefficients are provided. $^a = p < 0.05$, $^b = p < 0.01$, and $^c = p < 0.001$.

	Model	Sex	HS GPA	SAT/ACT	SE	TA	R^2
In-Person	Model 7a	-0.13^b	0.30^c	0.42^c	0.18^c	0.03^{ns}	0.39
	Model 7b	-0.14^c	0.30^b	0.43^c	0.19^c		0.39
	Model 7c	-0.16^c	0.31^c	0.42^c		0.11^b	0.37
	Model 7d	-0.19^c	0.31^c	0.45^c			0.36
Virtual	Model 8a	0.05^{ns}	0.23^c	0.29^c	0.24^b	-0.02^{ns}	0.25
	Model 8b	0.06^{ns}	0.23^c	0.29^c	0.24^c		0.25
	Model 8c	0.00^{ns}	0.25^c	0.35^c			0.21

Table 64: Physics 2 high-stakes assessment scores predicted by student sex, HS GPA, SAT/ACT Math scores, as well as pre self-efficacy and test anxiety. Standardized regression (β) coefficients are provided. For virtual classes, $N = 216$ and for in-person classes, $N = 426$. $^{ns} = p \geq 0.05$, $^{a} = p < 0.05$, $^{b} = p < 0.01$, and $^{c} = p < 0.001$.

	Model	Sex	HS GPA	SAT/ACT	SE	TA	R^2
In-Person	Model 9a	-0.05^{ns}	0.16^{a}	0.40^{c}	0.33^{c}	0.00^{ns}	0.34
	Model 9b	-0.09^{ns}	0.15^{a}	0.43^{c}	0.33^{c}		0.35
	Model 9c	-0.06^{ns}	0.16^{a}	0.40^{c}		0.15^{a}	0.37
	Model 9d	-0.12^{ns}	0.13^{ns}	0.48^{c}			0.25
Virtual	Model 10a	-0.08^{ns}	0.23^{c}	0.33^{c}	0.23^{b}	-0.07^{ns}	0.24
	Model 10b	-0.07^{ns}	0.23^{c}	0.32^{c}	0.19^{b}		0.24
	Model 10c	-0.11^{ns}	0.24^{c}	0.37^{c}			0.21

D.1 Mediation Models

A mediation model for physics 1 in-person classes can be seen in Figures 5a and 5b. Figure 5a shows that test anxiety is statistically significant when predicting high stakes assessment outcomes on its own. However, Figure 5b shows that if test anxiety is used to predict self-efficacy and high-stakes assessment outcomes separately, test anxiety predicts self-efficacy but not high-stakes grades. For virtual physics 1 classes, the average causal mediation effect was 0.24 ($p < 0.001$), with a confidence interval of [0.14, 0.34]. The average direct effect was 0.08, ($p = 0.260$) and the total direct effect was 0.32 ($p < 0.001$).

Appendix E. Original mindset survey items for an introductory physics course for engineering and physical science majors

Table 65: Original Survey Items. Italicized items were removed during validation.

	Construct name or Item
	My Growth ($\alpha = 0.84$)
1.	I can become even better at solving physics problems through hard work
2.	I am capable of really understanding physics if I work hard
3.	I can change my intelligence in physics quite a lot by working hard
4.	*Struggling with difficult physics problems would help me develop mastery in physics*
	My Ability ($\alpha = 0.84$)
5.	Even if I were to spend a lot of time working on difficult physics problems, I cannot develop my intelligence in physics further
6.	*If I were to often make mistakes on physics assignments and exams, I would think that maybe I'm just not smart enough to excel in physics.*
7.	I won't get better at physics if I try harder
8.	*I will always be as good at physics as I was in high school.*
9.	*I will always get the same physics grade whether I try or not.*
10.	I could never excel in physics because I do not have what it takes to be a physics person
11.	I could never become really good at physics even if I were to work hard because I don't have natural ability
	Others' Growth ($\alpha_{pre} = 0.84$)
7.	People can change their intelligence in physics quite a lot by working hard
8.	If people were to spend a lot of time working on difficult physics problems, they could develop their intelligence in physics quite a bit
9.	People can become good at solving physics problems through hard work
	Others' Ability ($\alpha_{pre} = 0.68$)
10.	Only a few specially qualified people are capable of really understanding physics
11.	To really excel in physics, people need to have a natural ability in physics
12.	If a student were to often make mistakes on physics assignments and exams, I would think that maybe they are just not smart enough to excel in physics

Appendix F. Invariance Testing of Intelligence Mindset Constructs

To determine if men and women could be included in the same factor analysis, we conducted tests for both strong and weak measurement invariance. First, we tested for weak measurement invariance: that is, does each survey item have similar factor loadings for men and women? To test for this, we compared the model fits of a free model (in which all factor loadings and intercepts are freely varying for each gender/sex group) and a metric model (in which the factor loadings are fixed to equity across gender/sex groups, but the intercepts were allowed to freely vary).

This free model (CLI = 0.95, TLI=0.93, RMSEA = 0.076, SRMR=0.056) and the metric model (CLI = 0.94, TLI = 0.93, RMSEA = 0.074, SRMR = 0.058) were not statistically significantly different ($\Delta\chi^2 = 15.2$, Δd.f. $= 9$, $p = 0.084$). Thus, we assumed weak invariance.

Next, we tested for strong measurement invariance: that is, does each survey item have similar factor loadings and intercept for men and women? We compared the model fits of a free model (in which all factor loadings and intercepts are freely varying for each gender group) and a scalar model (in which both the factor loadings and intercepts are fixed to equity across gender/sex groups). The chi-squared difference test between the free model and the scalar model (CLI = 0.94, TLI=0.94, RMSEA = 0.071, SRMR=0.059) was also nonsignificant ($\Delta\chi^2 = 23.0$, Δd.f. $= 18$, $p = 0.191$). The chi-squared difference test between the metric model and the scalar model was also nonsignificant ($\Delta\chi^2 = 7.74$, Δd.f. $= 9$, $p = 0.561$). Because of the non-significant chi-squared tests, we also assumed strong invariance, so we included both men and women in our factor analysis

Appendix G. Comparison of Students who Took the Mindset Pre-Survey Versus Pre- and Post-Surveys

Table 66: Comparison of sample size, gender/sex distribution, prior academic preparation and grades between students who only took the pre-survey, those who took both surveys, and all students enrolled in the class. Additionally, we use t-tests of mindset survey responses between students who took only the pre survey and those who took both. Numbers in parentheses are standard deviations.

	All Students	Pretest Only	Pre and Post	t-test	p
N	644	300	197		
% Women	36%	40%	36%		
HS GPA	4.12 (0.41)	4.14 (0.42)	4.19 (0.35)	-1.38	0.167
SAT Math	705 (62)	705 (64)	707 (56)	-0.19	0.852
Grade	2.62 (0.98)	2.60 (0.97)	2.73 (0.97)	-1.94	0.125
My Growth Pre		3.58 (0.49)	3.65 (0.44)	-1.66	0.098
My Ability Pre		3.40 (0.48)	3.46 (0.45)	-1.55	0.130
Others' Growth Pre		3.48 (0.50)	3.52 (0.48)	-0.94	0.346
Others' Ability Pre		3.15 (0.57)	3.17 (0.50)	-0.50	0.614

Appendix H. Original mindset survey items for an introductory physics course for bioscience majors

Table 67: Original Survey Items. Italicized items were removed from analysis during validation. Item 10 was removed during interviews because students did not interpret the question as intended and the rest were removed during statistical survey validation.

	Item
	My Growth
1.	I can become even better at solving physics problems through hard work
2.	I am capable of really understanding physics if I work hard
3.	I can change my intelligence in physics quite a lot by working hard
4.	*Struggling with difficult physics problems would help me develop mastery in physics*
	My Ability
5.	I won't get better at physics if I try harder
6.	I could never excel in physics because I do not have what it takes to be a physics person
7.	I could never become really good at physics even if I were to work hard because I don't have natural ability
8.	*If I were to often make mistakes on physics assignments and exams, I would think that maybe I'm just not smart enough to excel in physics.*
9.	*I won't get better at physics if I try harder*
10.	*I will always be as good at physics as I was in high school*
11.	*I will always get the same physics grade whether I try or not*
	Others' Growth
12.	People can change their intelligence in physics quite a lot by working hard
13.	If people were to spend a lot of time working on difficult physics problems, they could develop their intelligence in physics quite a bit
14.	People can become good at solving physics problems through hard work
15.	*If people were to persist in struggling with difficult physics problems, they would develop mastery in physics*
	Others' Ability
16.	Only a few specially qualified people are capable of really understanding physics
17.	To really excel in physics, people need to have a natural ability in physics
18.	If a student were to often make mistakes on physics assignments and exams, I would think that maybe they are just not smart enough to excel in physics
19.	*If people really have to struggle to solve physics problems, that means they are just not physics people.*

Appendix I. Predicting introductory physics for bioscience majors course grades with categorical mindset components

Table 68: Unstandardized regression coefficients for each average mindset categorical component. SAT/ACT math (“SAT/ACT” scores have been divided by 10 (such that they are on a 20-80). Medium (“Med”) scores are ≥ 2.5, and high scores are ≥ 3.5. The gender/sex (“Gen”) variable was coded such that women = 1 and men = 0. An “×” indicates an interaction term between two variables.
$^{*} = p < 0.05$, $^{**} = p < 0.01$, and $^{***} = p < 0.001$.

	My				Others'		
	Ability		Growth		Ability		Growth
Variable	Model 1	Model 2	Model 1	Model 2	Model 1	Model 2	Model 1
Gender	-0.22^{**}	-0.03	-0.25^{***}	-0.02	-0.26^{**}	-0.52^{*}	-0.29^{***}
HS GPA	0.62^{***}	0.62^{***}	0.63^{***}	0.63^{***}	0.66^{***}	0.66^{***}	0.65^{***}
SAT/ACT	0.05^{***}	0.05^{***}	0.04^{***}	0.04^{***}	0.05^{***}	0.05^{***}	0.05^{***}
Med	0.28^{**}	0.48	0.36	0.63	0.19^{*}	-0.05	-0.11
High	0.47^{***}	0.61	0.59^{**}	0.82	0.25^{*}	0.06	-0.04
Gen × Med		-0.23		-0.31		0.30	
Gen × High		-0.13		-0.23		0.22	
Adjusted R^2	0.39	0.39	0.40	0.39	0.38	0.38	0.3

Appendix J. Majors, courses, grades, and AGAs of Engineering majors taking introductory physics before, during, and after remote instruction

Table 69: Mean and standard deviation (SD) of average grades, as well as number of students (N) for each course of interest before the COVID-19 Pandemic, during remote classes due to COVID-19, and after the return to in-person instruction. Engineering Communication was a class that was required for students stating Spring 2020 and was not available during pre-remote instruction.

	Pre-Remote			Remote			Post-Remote		
Course	N	Mean	SD	N	Mean	SD	N	Mean	SD
Physics 1	949	2.63	0.830	336	2.98	0.70	531	2.56	1.04
Physics 2	614	2.73	0.800	244	2.97	0.78	307	2.75	0.79
Chemistry 1	652	2.48	0.990	246	2.48	0.92	363	2.24	1.18
Chemistry 2	440	2.36	0.840	205	2.49	0.85	287	2.24	1.02
Calculus 1	581	2.89	1.000	242	2.75	0.90	465	2.30	1.26
Calculus 2	604	2.59	1.100	211	3.02	0.74	323	2.57	1.18
Engineering Analysis	851	3.39	0.590	344	3.49	0.64	478	3.32	0.78
Engineering Computing	670	3.20	0.840	290	3.13	0.81	415	3.01	0.96
Composition Seminar	442	3.51	0.720	361	3.54	0.56	511	3.40	0.83
Engineering Communication				295	3.8	0.35	410	3.60	0.43

Table 70: Mean and standard deviation (SD) of average grade anomalies (AGA), as well as number of students (N) for each course of interest before the COVID-19 Pandemic, during remote classes due to COVID-19, and after the return to in-person instruction.. Engineering Communication was a class that was required for students stating Spring 2020 and was not available during pre-remote instruction.

	Pre-Remote			Remote			Post-Remote		
Course	N	Mean	SD	N	Mean	SD	N	Mean	SD
Physics 1	949	-0.64	1.01	336	-0.20	0.64	531	-0.62	1.26
Physics 2	614	-0.48	0.60	244	-0.32	0.64	307	-0.48	0.55
Chemistry 1	652	-0.79	1.07	246	-0.75	0.74	363	-0.96	1.11
Chemistry 2	440	-0.73	0.57	205	-0.70	0.60	287	-0.83	0.67
Calculus 1	581	-0.26	1.44	242	-0.48	1.05	465	-0.86	1.78
Calculus 2	604	-0.55	1.00	211	-0.21	0.85	323	-0.55	0.99
Engineering Analysis	851	0.36	0.57	344	0.44	0.46	478	0.43	0.66
Engineering Computing	670	0.18	0.75	290	0.00	0.60	415	0.01	0.73
Composition Seminar	442	0.43	0.77	361	0.48	0.57	511	0.49	0.88
Engineering Communication				295	0.73	0.58	410	0.65	0.66

Table 71: Average grade anomalies (AGAs), grades, and between-gender effect sizes for each course of interest in the four semesters before the COVID-19 Pandemic. Cohen's d is positive if men had higher grades or AGAs than women in a course. $^{\gamma} = p < 0.05$, $^{\beta} = p < 0.01$, and $^{\alpha} = p < 0.001$.

	Women					Men						
		AGA		Grade			AGA		Grade		Cohen's d	
Course	N	Mean	SD	Mean	SD	N	Mean	SD	Mean	SD	AGA	Grade
Physics 1	376	-0.84	0.94	2.49	0.77	573	-0.51	1.03	2.71	0.86	0.33^{α}	0.27^{α}
Physics 2	198	-0.67	0.55	2.59	0.74	416	-0.39	0.60	2.80	0.82	0.48^{α}	0.27^{β}
Chem 1	234	-0.83	1.17	2.50	1.00	418	-0.77	1.01	2.47	0.99	0.05	-0.03
Chem 2	148	-0.74	0.55	2.39	0.77	292	-0.73	0.59	2.34	0.87	0.02	-0.06
Calculus 1	216	-0.22	1.45	2.97	0.99	365	-0.28	1.43	2.85	1.00	-0.04	-0.12
Calculus 2	209	-0.55	1.06	2.61	1.16	395	-0.54	0.98	2.57	1.07	-0.01	-0.03
Analysis	320	0.36	0.53	3.43	0.57	531	0.36	0.60	3.37	0.6	0.01	0.10
Computing	231	0.10	0.79	3.15	0.92	439	0.22	0.73	3.23	0.80	0.15	0.10
Seminar	161	0.59	0.79	3.68	0.58	281	0.34	0.74	3.42	0.77	0.33^{β}	0.36^{α}

Table 72: Average grade anomalies (AGAs), grades, and between-gender effect sizes for each course of interest in the two semesters of remote instruction due to the COVID-19 Pandemic. We abbreviate the following course names: Engineering Analysis (Analysis), Engineering Computing (Computing), Composition Smeinar (Seminar), and Engineering Communication (Comm). Cohen's d is positive if men had higher grades or AGAs than women in a course.
$^{\gamma} = p < 0.05$, $^{\beta} = p < 0.01$, and $^{\alpha} = p < 0.001$.

	Women					Men						
		AGA		Grade			AGA		Grade		Cohen's d	
Course	N	Mean	SD	Mean	SD	N	Mean	SD	Mean	SD	AGA	Grade
Physics 1	127	-0.32	0.64	2.94	0.71	209	-0.12	0.63	3.00	0.7	0.32^{γ}	0.09
Physics 2	87	-0.28	0.54	3.04	0.73	157	-0.34	0.69	2.93	0.81	-0.09	-0.15
Chemistry 1	79	-0.74	0.74	2.53	0.91	167	-0.75	0.73	2.46	0.92	0.01	0.07
Chemistry 2	70	-0.60	0.52	2.63	0.80	135	-0.75	0.63	2.41	0.87	0.25	0.26
Calculus 1	94	-0.56	1.11	2.74	0.82	148	-0.43	1.02	2.76	0.95	0.12	0.01
Calculus 2	78	-0.31	0.51	2.96	0.68	133	-0.15	1.00	3.05	0.78	0.19	0.11
Analysis	132	0.42	0.48	3.53	0.65	212	0.45	0.45	3.47	0.64	0.08	-0.09
Computing	103	-0.06	0.57	3.10	0.78	187	0.03	0.62	3.15	0.82	0.15	0.06
Seminar	138	0.57	0.60	3.67	0.53	223	0.43	0.53	3.45	0.55	-0.25^{γ}	-0.39^{α}
Comm	104	0.75	0.54	3.83	0.30	191	0.73	0.60	3.78	0.37	-0.03	-0.17

Table 73: Average grade anomalies (AGAs), grades, and between-gender effect sizes for each course of interest in the two semesters of in-person instruction after remote classes due to COVID-19. We abbreviate the following course names: Engineering Analysis (Analysis), Engineering Computing (Computing), Composition Smeinar (Seminar), and Engineering Communication (Communication). Cohen's d is positive if men had higher grades or AGAs than women in a course. $^{\gamma} = p < 0.05$, $^{\beta} = p < 0.01$, and $^{\alpha} = p < 0.001$.

	Women					Men						
		AGA		Grade			AGA		Grade		Cohen's d	
Course	N	Mean	SD	Mean	SD	N	Mean	SD	Mean	SD	AGA	Grade
Physics 1	180	-0.95	1.55	2.36	1.03	351	-0.45	1.05	2.66	1.03	0.40^{α}	0.29^{β}
Physics 2	88	-0.61	0.50	2.72	0.74	219	-0.42	0.56	2.76	0.81	0.35^{β}	0.05
Chemistry 1	106	-1.11	1.41	2.23	1.25	257	-0.91	0.96	2.25	1.15	0.18	0.01
Chemistry 2	82	-0.81	0.62	2.27	0.94	205	-0.83	0.69	2.24	1.05	-0.04	-0.03
Calculus 1	155	-1.21	2.31	2.21	1.33	310	-0.68	1.42	2.35	1.22	0.30^{γ}	0.11
Calculus 2	92	-0.41	0.83	2.71	1.05	231	-0.61	1.05	2.51	1.22	-0.20	-0.17
Analysis	156	0.40	0.72	3.28	0.82	322	0.45	0.62	3.33	0.76	0.08	0.07
Computing	126	-0.14	0.74	2.89	0.98	289	0.08	0.71	3.06	0.94	0.29^{γ}	0.17
Seminar	164	0.62	0.81	3.48	0.81	347	0.43	0.90	3.36	0.83	-0.21^{γ}	-0.14
Comm	126	0.67	0.63	3.62	0.43	284	0.64	0.67	3.59	0.43	-0.05	-0.08

Appendix K. Majors and Courses for Bioscience Majors taking introductory physics before, during, and after remote instruction

Table 74: Course Requirements by Major. R designates a required course, O designates a course that can be taken for elective credit in the major, and no letter designates a course that does not fulfill any credits for the major. The following terms are abbreviated: Computational (Comp), Ecology and Evolution (E&E), Rehabilitation (Rehab), Calculus (Calc), Chemistry (Chem), Genetics (Gen), Organic Chemitry (Organic), Human Physiology (HP), and Biochemistry (BC).

		Biology		Chem			Organic				Physics	
Major	Calc 1	1	2	1	2	Gen	1	2	HP	BC	1	2
Biology	R	R	R	R	R	R	R	R	O	R	R	R
Comp Biology	R	R	R	R	R	R	R			R	O	
E&E	R	R	R	R	R	R	R	R		R	R	R
Microbiology	R	R	R	R	R	R	R	R	O	R	R	R
Molecular Biology	R	R	R	R	R	R	R	R			R	R
Neuroscience	R	R	R	R	R		R	R	R	R	R	R
Pharmacy	R	R	R	R	R		R	R		R	O	O
Rehab Science		R		R					O		R	

Table 75: List of courses studied, the department that offers them, and the percentage of students in our sample who take each course in a given year. For example, 61% of students take calculus during their first year of university, and 11% of students take calculus during their second year. The year in which students take the course most often has its percentage of students in bold.

Course	Course Type	1^{st}	2^{ed}	3^{ed}	4^{th}	$\geq 5^{th}$
Calculus 1	Pre-Remote	**68**	21	6	4	1
	Remote	**60**	27	5	4	4
	Post-Remote	**39**	35	19	6	2
Biology 1	Pre-Remote	**91**	8	1	0	0
	Remote	**80**	17	2	1	0
	Post-Remote	**67**	21	8	2	1
Biology 2	Pre-Remote	**72**	23	4	1	0
	Remote	**60**	33	6	1	1
	Post-Remote	41	**44**	11	4	0
Chemistry 1	Pre-Remote	**93**	6	0	0	0
	Remote	**86**	11	2	1	0
	Post-Remote	**68**	22	8	1	1
Chemistry 2	Pre-Remote	**77**	19	3	1	0
	Remote	**65**	26	6	1	1
	Post-Remote	**45**	39	12	3	1
Organic Chemistry 1	Pre-Remote	5	**81**	11	2	1
	Remote	7	**83**	7	2	0
	Post-Remote	8	**69**	18	3	1
Organic Chemistry 2	Pre-Remote	2	**63**	28	5	1
	Remote	3	**70**	18	6	3
	Post-Remote	3	**54**	32	8	3
Genetics	Pre-Remote	4	**54**	32	9	1
	Remote	2	**49**	32	12	4
	Post-Remote	4	30	**46**	16	4
Physics 1	Pre-Remote	13	30	**51**	5	1
	Remote	14	32	**45**	7	2
	Post-Remote	15	32	**47**	5	2
Physics 2	Pre-Remote	2	23	**59**	12	2
	Remote	4	18	**60**	14	4
	Post-Remote	3	21	**61**	12	3
Human Physiology	Pre-Remote	2	20	**61**	15	2
	Remote	1	12	**51**	30	6
	Post-Remote	1	11	**54**	31	4
Biochemistry	Pre-Remote	1	8	**60**	27	5
	Remote	1	8	**60**	24	7
	Post-Remote	0	8	**64**	21	8

Appendix L. Grades and Grade Anomalies for Bioscience Majors taking introductory physics before, during, and after remote instruction

Table 76: Course grades and number of students before the COVID-19 Pandemic, during remote classes due to COVID-19, and after the return to in-person instruction.

	Pre-Remote			Remote			Post-Remote		
Course	N	Mean	SD	N	Mean	SD	N	Mean	SD
Calculus 1	812	2.46	1.22	448	2.77	0.96	293	2.07	1.45
Biology 1	1648	2.87	1.01	511	3.04	0.86	217	2.40	1.17
Biology 2	1421	3.10	0.77	549	3.27	0.80	212	2.77	0.91
Chemistry 1	1839	2.85	0.91	598	3.05	0.77	261	2.44	1.05
Chemistry 2	1567	2.90	0.88	576	2.80	0.85	312	2.25	1.09
O Chemistry 1	1634	2.50	1.10	800	2.79	0.93	537	2.28	1.14
O Chemistry 2	1102	2.59	1.12	654	2.72	1.02	421	2.22	1.25
Genetics	758	2.91	1.06	439	3.40	0.81	363	2.74	1.07
Physics 1	1158	3.07	0.80	709	3.27	0.80	881	2.77	0.96
Physics 2	741	3.16	0.82	647	3.38	0.74	636	3.13	0.95
Human Physiology	905	3.20	0.89	710	3.37	0.76	809	3.00	0.98
Biochemistry	524	2.50	1.08	583	3.36	0.81	575	2.79	1.25

Table 77: Course grade anomalies before the COVID-19 Pandemic, during remote classes due to COVID-19, and after the return to in-person instruction. Mean and standard deviation (SD) of average grade anomalies (AGA), as well as number of students (N) for each course of interest.

	Pre-Remote			Remote			Post-Remote		
Course	N	Mean	SD	N	Mean	SD	N	Mean	SD
Calculus 1	812	-0.85	1.72	448	-0.59	0.91	293	-1.12	1.58
Biology 1	1648	-0.65	1.09	511	-0.43	0.73	217	-0.69	1.08
Biology 2	1421	-0.35	0.64	549	-0.13	0.64	212	-0.51	0.68
Chemistry 1	1839	-0.73	1.40	598	-0.36	0.69	261	-0.73	1.02
Chemistry 2	1567	-0.56	0.80	576	-0.65	0.68	312	-0.99	0.88
Organic Chemistry 1	1634	-0.98	0.98	800	-0.77	0.76	537	-1.10	0.97
Organic Chemistry 2	1102	-0.90	0.95	654	-0.85	0.84	421	-1.19	1.06
Genetics	758	-0.55	0.88	439	-0.13	0.62	363	-0.64	0.80
Physics 1	1158	-0.36	0.62	709	-0.22	0.60	881	-0.64	0.79
Physics 2	741	-0.31	0.63	647	-0.17	0.59	636	-0.37	0.75
Human Physiology	905	-0.32	0.74	710	-0.14	0.63	809	-0.56	0.81
Biochemistry	524	-0.96	0.90	583	-0.14	0.67	575	-0.70	1.01

Table 78: Average grade anomalies (AGAs), grades, and between-gender effect sizes for each course of interest in the four semesters beofore the COVID-19 Pandemic. Organic Chemistry (O Chem) and Human Physiology (Human Phys) are abbreviated. Cohen's d is positive if men had higher grades or AGAs than women in a course. $^{\gamma} = p < 0.05$, $^{\beta} = p < 0.01$, and $^{\alpha} = p < 0.001$.

	Women					Men						
		AGA		Grade			AGA		Grade		Cohen's d	
Course	N	Mean	SD	Mean	SD	N	Mean	SD	Mean	SD	AGA	Grade
Calculus 1	379	-0.78	1.49	2.60	1.11	433	-0.92	1.91	2.35	1.3	0.08	0.21$^{\beta}$
Biology 1	1079	-0.70	1.17	2.86	1.02	569	-0.56	0.92	2.88	1.00	0.13$^{\gamma}$	0.02
Biology 2	902	-0.38	0.65	3.10	0.77	519	-0.3	0.64	3.11	0.79	0.12$^{\gamma}$	0.02
Chemistry 1	1160	-0.79	1.38	2.86	0.87	679	-0.61	1.41	2.83	0.97	0.13$^{\gamma}$	0.04
Chemistry 2	970	-0.58	0.77	2.94	0.86	597	-0.52	0.86	2.85	0.91	0.07	0.09
O Chem 1	1051	-1.04	0.99	2.47	1.12	583	-0.87	0.94	2.57	1.07	0.18$^{\alpha}$	0.09
O Chem 2	682	-0.95	0.92	2.56	1.08	420	-0.82	0.98	2.64	1.17	0.13$^{\gamma}$	0.07
Genetics	453	-0.48	0.78	3.02	0.96	305	-0.67	1.01	2.74	1.18	0.22$^{\gamma}$	0.26$^{\beta}$
Physics 1	720	-0.42	0.61	3.05	0.77	438	-0.26	0.64	3.12	0.84	0.26$^{\alpha}$	0.09
Physics 2	449	-0.33	0.61	3.17	0.78	292	-0.28	0.66	3.15	0.88	0.08	0.02
Human Phys	571	-0.36	0.76	3.18	0.90	334	-0.25	0.69	3.22	0.87	0.14$^{\gamma}$	0.05
Biochemistry	305	-0.93	0.88	2.57	1.03	219	-1.00	0.92	2.41	1.15	0.09	0.15

Table 79: Average grade anomalies (AGAs), grades, and between-gender effect sizes for each course of interest in the two semesters of remote instruction due to the COVID-19 Pandemic. Organic Chemistry (O Chem) and Human Physiology (Human Phys) are abbreviated. Cohen's d is positive if men had higher grades or AGAs than women in a course. $^{\gamma} = p < 0.05$, $^{\beta} = p < 0.01$, and $^{\alpha} = p < 0.001$.

	Women					Men						
		AGA		Grade			AGA		Grade		Cohen's d	
Course	N	Mean	SD	Mean	SD	N	Mean	SD	Mean	SD	AGA	Grade
Calculus 1	226	-0.70	0.95	2.69	0.98	222	-0.47	0.85	2.86	0.93	0.26*	0.17
Biology 1	326	-0.50	0.74	3.04	0.83	185	-0.32	0.70	3.03	0.91	0.25^{γ}	0.01
Biology 2	363	-0.16	0.65	3.26	0.80	186	-0.07	0.62	3.29	0.79	0.14	0.03
Chemistry 1	374	-0.37	0.70	3.1	0.74	224	-0.35	0.69	2.97	0.83	0.04	0.17
Chemistry 2	359	-0.70	0.69	2.80	0.83	217	-0.55	0.65	2.80	0.90	0.23^{γ}	0.00
O Chem 1	530	-0.81	0.79	2.77	0.95	270	-0.69	0.70	2.83	0.88	0.16	0.06
O Chem 2	441	-0.90	0.81	2.71	0.97	213	-0.75	0.90	2.73	1.12	0.18	0.02
Genetics	267	-0.12	0.62	3.42	0.80	172	-0.14	0.63	3.35	0.84	0.04	0.08
Physics 1	414	-0.29	0.60	3.22	0.78	295	-0.13	0.58	3.35	0.81	0.26^{β}	0.16^{γ}
Physics 2	394	-0.22	0.63	3.37	0.71	253	-0.08	0.51	3.40	0.79	0.23^{γ}	0.05
Human Phys	480	-0.14	0.61	3.34	0.75	230	-0.12	0.67	3.44	0.78	0.03	0.13
Biochemistry	349	-0.19	0.67	3.37	0.81	234	-0.08	0.68	3.33	0.82	0.16	0.05

Table 80: Average grade anomalies (AGAs), grades, and between-gender effect sizes for each course of interest in the two semesters of in-person instruction after remote classes due to COVID-19. Cohen's d is positive if men had higher grades or AGAs than women in a course. Organic Chemistry (O Chem) and Human Physiology (Human Phys) are abbreviated. $^{\gamma} = p < 0.05$, $^{\beta} = p < 0.01$, and $^{\alpha} = p < 0.001$.

	Women					Men						
		AGA		Grade			AGA		Grade		Cohen's d	
Course	N	Mean	SD	Mean	SD	N	Mean	SD	Mean	SD	AGA	Grade
Calculus 1	131	-1.37	1.57	1.92	1.45	162	-0.92	1.55	2.19	1.44	0.29^{γ}	0.19
Biology 1	141	-0.87	1.15	2.28	1.22	76	-0.36	0.87	2.62	1.03	0.48^{α}	0.30^{γ}
Biology 2	121	-0.57	0.65	2.77	0.85	91	-0.44	0.72	2.77	0.98	0.20	0.00
Chemistry 1	134	-0.84	1.13	2.44	1.09	127	-0.61	0.88	2.44	1.02	0.23	0.00
Chemistry 2	174	-1.13	0.91	2.17	1.08	138	-0.81	0.82	2.36	1.09	0.37	0.18
O Chem 1	340	-1.3	0.91	2.11	1.09	197	-0.76	0.97	2.58	1.16	0.58^{α}	0.42^{α}
O Chem 2	249	-1.37	1.02	2.07	1.22	172	-0.93	1.08	2.43	1.28	0.43^{α}	0.30^{β}
Genetics	225	-0.72	0.80	2.7	1.02	138	-0.50	0.79	2.8	1.14	0.27^{γ}	0.10
Physics 1	590	-0.72	0.79	2.74	0.95	291	-0.48	0.76	2.85	0.98	0.31^{α}	0.12
Physics 2	384	-0.40	0.76	3.12	0.93	252	-0.33	0.73	3.15	0.98	0.10	0.03
Human Phys	527	-0.58	0.80	2.98	0.96	282	-0.51	0.82	3.04	1.00	0.10	0.06
Biochemistry	376	-0.74	1.02	2.78	1.24	199	-0.64	0.98	2.81	1.26	0.10	0.03

Bibliography

[1] Women, Minorities, and Persons with Disabilities in Science and Engineering: 2019, 2020. https://ncses.nsf.gov/pubs/nsf19304/digest/field-of-degree-women.

[2] Nina Abramzon, Patrice Benson, Edmund Bertschinger, Susan Blessing, Geraldine L Cochran, Anne Cox, Beth Cunningham, Jessica Galbraith-Frew, Jolene Johnson, Leslie Kerby, Elaine Lalanne, Christine O'Donnell, Sara Petty, Sujatha Sampath, Susan Seestrom, Chandralekha Singh, Cherrill Spencer, Kathryne Sparks Woodle, and Sherry Yennello. Women in physics in the United States: Recruitment and retention. *AIP Conference Proceedings*, 1697(1):060045, 2015.

[3] Danny Doucette and Chandralekha Singh. Why are there so few women in physics? Reflections on the experiences of two women. *The Physics Teacher*, 58(5):297–300, 2020.

[4] H B Gonzalez and J J Kuenzi. Science, Technology, Engineering, and Mathematics (STEM) Education: A Primer. Report CRS Report No. R42530, Library of Congress Congressional Research Service, 2012.

[5] Jennifer Blue, Adrienne Traxler, and Ximena Cid. Gender matters. *Physics Today*, 71(3):40–46, 2018.

[6] Therese Bouffard-Bouchard, Sophie Parent, and Serge Larivee. Influence of self-efficacy on self-regulation and performance among junior and senior high-school age students. *International Journal of Behavioral Development*, 14(2):153–164, 1991.

[7] Paul R Pintrich and Elisabeth V. De Groot. Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology*, 82(1):33–40, 1990.

[8] Barry J Zimmerman. Self-efficacy: An essential motive to learn. *Contemporary Educational Psychology*, 25(1):82–91, 2000.

[9] Z Yasemin Kalender, Emily Marshman, Christian D Schunn, Timothy J Nokes-Malach, and Chandralekha Singh. Why female science, technology, engineering, and mathematics majors do not identify with physics: They do not think others see them that way. *Physical Review Physics Education Research*, 15(2):020148, 2019.

[10] Yangqiuting Li and Chandralekha Singh. Effect of gender, self-efficacy, and interest on perception of the learning environment and outcomes in calculus-based introductory physics courses. *Physical Review Physics Education Research*, 17(1):010143, 2021.

[11] Jayson M Nissen and Jonathan T Shemwell. Gender, experience, and self-efficacy in introductory physics. *Physical Review Physics Education Research*, 12(2):020105, 2016.

[12] Vashti Sawtelle, Eric Brewe, and Laird H Kramer. Exploring the relationship between self-efficacy and retention in introductory physics. *Journal of Research in Science Teaching*, 49(9):1096–1121, 2012.

[13] Chandralekha Singh and Alysa Malespina. Test anxiety, self-efficacy, and gender: A quest for equitable assessment practices in physics. In *2021 Physics Education Research Conference Proceedings*, 2021. https://doi.org/10.1119/perc.2021.pr.Singh.

[14] Alysa Malespina and Chandralekha Singh. Impact of grade penalty in first-year foundational science courses on female engineering majors. *International Journal of Engineering Education*, 38(4):1021–1031. https://www.ijee.ie/latestissues/Vol38-4/13$_{i}jee4218.pdf, year = 2022$.

[15] Alysa Malespina and Chandralekha Singh. Gender differences in grades versus grade penalties: Are grade anomalies more detrimental for female physics majors? *Physical Review Physics Education Research*, 18:020127, 2022.

[16] Alysa Malespina, Christian D Schunn, and Chandralekha Singh. Whose ability and growth matter? Gender, mindset and performance in physics. *International Journal of STEM Education*, 9, 2022.

[17] Alysa Malespina, Christian Schunn, and Chandralekha Singh. To whom do students believe a growth mindset applies? In *2022 Physics Education Research Conference Proceedings*, 2022. https://doi.org/10.1119/perc.2022.pr.Malespina.

[18] Alysa Malespina and Chandralekha Singh. Gender differences in test anxiety and self-efficacy: Why instructors should emphasize low-stakes formative assessments in physics courses. *European Journal of Physics*, 43(3):035701, 2022.

[19] Alysa Malespina, Christian D Schunn, and Chandralekha Singh. Gender gaps in grades versus grade penalties: Why grade anomalies may be more detrimental for women aspiring for careers in biological sciences. *International Journal of STEM Education*, 10(13), 2023. https://doi.org/10.1186/s40594-023-00399-7.

[20] Alysa Malespina, Christian Schunn, and Chandralekha Singh. Bioscience students' internalized mindsets predict grades and reveal gender inequities in physics courses (accepted). *Physical Review Physics Education Research*, 2023.

[21] Alysa Malespina and Chandralekha Singh. Peer interaction, self-efficacy, and equity: Same gender groups are more beneficial than mixed gender groups for female students (accepted). *Journal of College Science Teaching*, 2023.

[22] Sonja Cwik and Chandralekha Singh. How perception of learning environment predicts male and female students' grades and motivational outcomes in algebra-based introductory physics courses. *Physical Review Physics Education Research*, 17:020143, 2021.

[23] Sonja Cwik and Chandralekha Singh. Damage caused by societal stereotypes: Women have lower physics self-efficacy controlling for grade even in courses

in which they outnumber men. *Physical Review Physics Education Research*, 17:020138, 2021.

[24] Sonja Cwik and Chandralekha Singh. Students' sense of belonging in introductory physics course for bioscience majors predicts their grade. *Physical Review Physics Education Research*, 18:010139, 2022.

[25] Sonja Cwik and Chandralekha Singh. Role of inclusiveness of learning environment in predicting students' outcomes in courses in which women are not underrepresented. *Journal of Higher Education Theory and Practice*, 22(17):176–189, 2022.

[26] Sonja Cwik and Chandralekha Singh. Longitudinal analysis of women and men's motivational beliefs in a two-semester introductory physics course sequence for students on the bioscience track. *Physical Review Physics Education Research*, 18(2):020111, 2022.

[27] Sonja Cwik and Chandralekha Singh. Not feeling recognized as a physics person by instructors and teaching assistants is correlated with female students' lower grades. *Physical Review Physics Education Research*, 18:010138, 2022.

[28] Yangqiuting Li, Kyle Whitcomb, and Chandralekha Singh. How perception of being recognized or not recognized by instructors as a "physics person" impacts male and female students' self-efficacy and performance. *The Physics Teacher*, 58(7):484–487, 2020.

[29] Yangqiuting Li and Chandralekha Singh. Do female and male students' physics motivational beliefs change in a two-semester introductory physics course sequence? *Physical Review Physics Education Research*, 18:010142, 2022.

[30] Yangqiuting Li and Chandralekha Singh. How engineering identity of first-year female and male engineering majors is predicted by their physics self-efficacy and identity. *International Journal of Engineering Education*, 38(3):1–15, 2022. https://www.ijee.ie/latestissues/Vol38-3/21_ijee4203.pdf.

[31] Yangqiuting Li and Chandralekha Singh. Inclusive learning environments can improve student learning and motivational beliefs. *Physical Review Physics Education Research*, 18:020147, 2022.

[32] Z Yasemin Kalender, Emily Marshman, Christian D Schunn, Timothy J Nokes-Malach, and Chandralekha Singh. Damage caused by women's lower self-efficacy on physics learning. *Physical Review Physics Education Research*, 16(1):010118, 2020.

[33] Z Yasemin Kalender, Emily Marshman, Christian D Schunn, Timothy J Nokes-Malach, and Chandralekha Singh. Framework for unpacking students' mindsets in physics by gender. *Physical Review Physics Education Research*, 18:010116, 2022.

[34] Danny Doucette, Russell Clark, and Chandralekha Singh. Hermione and the secretary: How gendered task division in introductory physics labs can disrupt equitable learning. *European Journal of Physics*, 41(3):035702, 2020.

[35] Danny Doucette and Chandralekha Singh. Views of female students who played the role of group leaders in introductory physics labs. *European Journal of Physics*, 42(3):035702, 2021.

[36] Danny Doucette and Chandralekha Singh. Share it, don't split it: Can equitable group work improve student outcomes? *The Physics Teacher*, 60:166–168, 2022.

[37] Danny Doucette and Chandralekha Singh. Making lab group work equitable and inclusive. *Journal of College Science Teaching*, 52(4):31–37, 2023. https://www.nsta.org/journal-college-science-teaching/journal-college-science-teaching-marchapril-2023/making-lab-group.

[38] Alexandru Maries, Nafis I Karim, and Chandralekha Singh. Is agreeing with a gender stereotype correlated with the performance of female students in introductory physics? *Physical Review Physics Education Research*, 14:020119, 2018.

[39] Alexandru Maries, Kyle Whitcomb, and Chandralekha Singh. Gender inequities throughout STEM. *Journal of College Science Teaching*, 51:27–36, 2022. https://www.nsta.org/journal-college-science-teaching/journal-college-science-teaching-januaryfebruary-2022/gender.

[40] Emily Marshman, Zeynep Y Kalender, Christian Schunn, Timothy Nokes-Malach, and Chandralekha Singh. A longitudinal analysis of students' motivational characteristics in introductory physics courses: Gender differences. *Canadian Journal of Physics*, 96(4):391–405, 2017.

[41] Emily M Marshman, Z Yasemin Kalender, Timothy Nokes-Malach, Christian Schunn, and Chandralekha Singh. Female students with A's have similar physics self-efficacy as male students with C's in introductory courses: A cause for alarm? *Physical Review Physics Education Research*, 14(2):020123, 2018.

[42] Emily Marshman, Zeynep Y Kalender, Christian Schunn, Timothy Nokes-Malach, and Chandralekha Singh. A longitudinal analysis of students' motivational characteristics in introductory physics courses: Gender differences. *Canadian Journal of Physics*, 96(4):391–405, 2018.

[43] Kyle M Whitcomb and Chandralekha Singh. For physics majors, gender differences in introductory physics do not inform future physics performance. *European Journal of Physics*, 41(6):065701, 2013.

[44] Kyle M Whitcomb, Z Yasemin Kalender, Timothy Nokes-Malach, Christian D Schunn, and Chandralekha Singh. Comparison of self-efficacy and performance of engineering undergraduate women and men. *International Journal of Engineering Education*, 34(4):1996–2004, 2020. https://www.ijee.ie/1atestissues/Vol36-6/24_ijee4004.pdf.

[45] Kyle M Whitcomb, Alexandru Maries, and Chandralekha Singh. Examining gender differences in a mechanical engineering and materials science curriculum. *International Journal of Engineering Education*, 37(5):1261–1273, 2021. https://www.ijee.ie/latestissues/Vol37-5/10_ijee4103.pd.

[46] Kyle M Whitcomb, Sonja Cwik, and Chandralekha Singh. Not all disadvantages are equal: Racial/ethnic minority students have largest disadvantage among demographic groups in both stem and non-stem gpa. *AERA Open*, 7:23328584211059823, 2021.

[47] Kyle M Whitcomb and Chandralekha Singh. Underrepresented minority students receive lower grades and have higher rates of attrition across STEM disciplines: A sign of inequity? *International Journal of Science Education*, 43(7):1054–1089, 2021.

[48] Kyle M Whitcomb, Danny Doucette, and Chandralekha Singh. Comparing major declaration, attrition, migration, and completion in physics with other STEM disciplines: A sign of inequitable physics culture? *Journal of Higher Education Theory and Practice*, 22(17):84–102, 2022.

[49] Kyle M Whitcomb, Alexandru Maries, and Chandralekha Singh. Progression in self-efficacy, interest, identity, sense of belonging, perceived recognition and effectiveness of peer interaction of physics majors and comparison with non-majors and Ph.D. students. *Research in Science Education*, 53(3):525–539, 2023.

[50] Lisabeth M Santana and Chandralekha Singh. Negative impacts of an unwelcoming physics environment on undergraduate women. In *Physics Education Research Conference 2021*, PER Conference, pages 377–383, Virtual Conference, August 4-5 2021. https://doi.org/10.1119/perc.2021.pr.Santana.

[51] Ming-Te Wang and Jessica Degol. Motivational pathways to STEM career choices: Using expectancy–value perspective to understand individual and gender differences in STEM fields. *Developmental Review*, 33(4):304–340, 2013.

[52] Kimberly Grau Talley and Araceli Martinez Ortiz. Women's interest development and motivations to persist as college students in STEM: A mixed methods analysis of views and voices from a Hispanic-serving

institution. *International Journal of STEM Education*, 4(5), 2013. https://doi.org/10.1186/s40594-017-0059-2.

[53] Elaine Seymour, Anne-Barrie Hunter, Heather Thiry, Timothy J Weston, Raquel P Harper, Dana G Holland, Andrew K Koch, and Brent M Drake. *Talking About Leaving Revisited: Persistence, Relocation, and Loss in Undergraduate STEM Education.* Springer International Publishing AG, Cham, Switzerland, 2019.

[54] Elaine Seymour, Nancy M Hewitt, and Cynthia M Friend. *Talking About Leaving: Why Undergraduates Leave the Sciences.* Westview Press, Boulder, CO, 1997.

[55] Ernesto Reuben, Paola Sapienza, and Luigi Zingales. How stereotypes impair women's careers in science. *Proceedings of the National Academy of Sciences - PNAS*, 111(12):4403–4408, 2014.

[56] Asia A Eaton, Jessica F Saunders, Ryan K Jacobson, and Keon West. How gender and race stereotypes impact the advancement of scholars in STEM: Professors' biased evaluations of physics and biology post-doctoral candidates. *Sex Roles*, 82(3-4):127–141, 2020.

[57] Lin Bian, Sarah-Jane Leslie, and Andrei Cimpian. Gender stereotypes about intellectual ability emerge early and influence children's interests. *Science*, 355(6323):389–391, 2017.

[58] Lin Bian, Sarah-Jane Leslie, and Andrei Cimpian. Evidence of bias against girls and women in contexts that emphasize intellectual ability. *American Psychologist*, 73(9):1139–1153, 2018.

[59] Sarah-Jane Leslie, Andrei Cimpian, Meredith Meyer, and Edward Freeland. Expectations of brilliance underlie gender distributions across academic disciplines. *Science*, 347(6219):262–265, 2015.

[60] Moshe Zeidner. *Test Anxiety: The State of the Art*. Springer, New York, New York, 1998.

[61] Chandralekha Singh. Problem solving and learning. *AIP Conference Proceedings*, 1140(1):183–197, 2009. https://doi.org/10.1063/1.3183522.

[62] Daniel Solomon, Victor Battistich, Dong-il Kim, and Marilyn Watson. Teacher practices associated with students' sense of the classroom as a community. *Social Psychology of Education*, 1(3):235–267, 1996.

[63] Albert Bandura. *Self-efficacy: The Exercise of Control*. Macmillan, 1997.

[64] Albert Bandura. On the functional properties of perceived self-efficacy revisited. *Journal of Management*, 38(1):9–44, 2012.

[65] Ann M L Cavallo, Wendell H Potter, and Michelle 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 Science and Mathematics*, 104(6):288–300, 2004.

[66] Anne Marie Porter and Rachel Ivie. *Women in Physics and Astronomy*. Statistical Research Center of the American Institute of Physics, College Park, MD, 2019. https://eric.ed.gov/?id=ED594227.

[67] Martha M Bleeker and Janis E. Jacobs. Achievement in math and science: Do mothers' beliefs matter 12 years later? *Journal of Educational Psychology*, 96(1):97–109, 2004.

[68] Carol S Dweck. *Mindset: The New Psychology of Success*. Random House, 2006.

[69] Katherine Muenks and David B Miele. Students' thinking about effort and ability: The role of developmental, contextual, and individual difference factors. *Review of Educational Research*, 87(4):707–735, 2017.

[70] David Scott Yeager and Carol S Dweck. Mindsets that promote resilience: When students believe that personal characteristics can be developed. *Educational Psychologist*, 47(4):302–314, 2012.

[71] Aneeta Rattan, Krishna Savani, Dolly Chugh, and Carol S Dweck. Leveraging mindsets to promote academic achievement: Policy recommendations. *Perspectives on Psychological Science*, 10(6):721–726, 2015.

[72] Carol S Dweck. *Is Math a Gift? Beliefs That Put Females at Risk*. American Psychological Association, 2007.

[73] Lisa B Limeri, Nathan T Carter, Jun Choe, Hannah G Harper, Hannah R Martin, Annaleigh Benton, and Erin L Dolan. Growing a growth mindset: Characterizing how and why undergraduate students' mindsets change. *International Journal of STEM Education*, 7(1):35, 2020.

[74] Catherine Good, Joshua Aronson, and Michael Inzlicht. Improving adolescents' standardized test performance: An intervention to reduce the effects of stereotype threat. *Journal of Applied Developmental Psychology*, 24(6):645–662, 2003.

[75] Lisa S Blackwell, Kali H Trzesniewski, and Carol Sorich Dweck. Implicit theories of intelligence predict achievement across an adolescent transition: A longitudinal study and an intervention. *Journal of Child development*, 78(1):246–263, 2007.

[76] A Rattan, K Savani, M Komarraju, M M Morrison, C Boggs, and N Ambady. Meta-lay theories of scientific potential drive underrepresented students' sense of belonging to science, technology, engineering, and mathematics (STEM). *Journal of Personality and Social Psychology*, 115(1):54–75, 2018.

[77] Claude M Steele and Joshua Aronson. Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology*, 69(5):797–811, 1995.

[78] Jacquelynne S Eccles and Allan Wigfield. From expectancy-value theory to situated expectancy-value theory: A developmental, social cognitive, and sociocultural perspective on motivation. *Contemporary Educational Psychology*, 61:101859, 2020.

[79] Janet T. Spence. *Achievement and Achievement Motives: Psychological and Sociological Approaches*. W.H. Freeman, San Francisco, 1983.

[80] Allan Wigfield and Jacquelynne S Eccles. *Development of Achievement Motivation*. Educational Psychology Series. Academic Press, San Diego, 2002.

[81] Shima Salehi, Eric Burkholder, G Peter Lepage, Steven Pollock, and Carl Wieman. Demographic gaps or preparation gaps?: The large impact of incoming preparation on performance of students in introductory physics. *Physical Review Physics Education Research*, 15:020114, 2019.

[82] Adrian Madsen, Sarah B McKagan, and Eleanor C Sayre. Gender gap on concept inventories in physics: What is consistent, what is inconsistent, and what factors influence the gap? *Physical Review Physics Education Research*, 9:020121, 2013.

[83] Philip M Sadler and Robert H Tai. Success in introductory college physics: The role of high school preparation. *Science Education*, 85(2):111–136, 2001.

[84] Simon Bates, Robyn Donnelly, Cait MacPhee, David Sands, Marion Birch, and Niels R Walet. Gender differences in conceptual understanding of newtonian mechanics: A UK cross-institution comparison. *European Journal of Physics*, 34(2):421–434, 2013.

[85] Alexandru Maries, Nafis I Karim, and Chandralekha Singh. Is agreeing with a gender stereotype correlated with the performance of female students in introductory physics? *Physical Review Physics Education Research*, 14(2):020119, 2018.

[86] Z Yasemin Kalender, Emily Marshman, Christian D Schunn, Timothy J Nokes-Malach, and Chandralekha Singh. Gendered patterns in the construction of physics identity from motivational factors. *Physical Review Physics Education Research*, 15(2):020119, 2019.

[87] Valerie Gibson, Lisa Jardine-Wright, and Elizabeth Bateman. An investigation into the impact of question structure on the performance of first year physics undergraduate students at the University of Cambridge. *European Journal of Physics*, 36(4):045014, 2015.

[88] Holly Hedgeland, Hillary Dawkins, and Sally Jordan. Investigating male bias in multiple choice questions: Contrasting formative and summative settings. *European Journal of Physics*, 39(5):055704, 2018.

[89] Hillary Dawkins, Holly Hedgeland, and Sally Jordan. Impact of scaffolding and question structure on the gender gap. *Physical Review Physics Education Research*, 13(2):020117, 2017.

[90] Melanie Good, Alexandru Maries, and Chandralekha Singh. Impact of traditional or evidence-based active-engagement instruction on introductory female and male students' attitudes and approaches to physics problem solving. *Physical Review Physics Education Research*, 15(2):020129, 2019.

[91] Nafis I Karim, Alexandru Maries, and Chandralekha Singh. Do evidence-based active-engagement courses reduce the gender gap in introductory physics? *European Journal of Physics*, 39(2):025701, 2018.

[92] Mercedes Lorenzo, Catherine H Crouch, and Eric Mazur. Reducing the gender gap in the physics classroom. *American Journal of Physics*, 74(2):118–122, 2006.

[93] Jared B Stang, Emily Altiere, Joss Ives, and Patrick J Dubois. Exploring the contributions of self-efficacy and test anxiety to gender differences in assessments. In *Physics Education Research Conference 2020*, PER Conference, pages 497–502, Virtual Conference, July 22-23 2020.

[94] Cissy J Ballen, Shima Salehi, and Sehoya Cotner. Exams disadvantage women in introductory biology. *PloS One*, 12(10):e0186419, 2017.

[95] Moshe Zeidner. *Test Anxiety: The State of the Art.* Springer, 1998.

[96] Martha M Bleeker and Janis E Jacobs. Achievement in math and science: Do mothers' beliefs matter 12 years later? *Journal of Educational Psychology*, 96(1):97–109, 2004.

[97] Janis E Jacobs and Jacquelynne S Eccles. The impact of mothers' gender-role stereotypic beliefs on mothers' and children's ability perceptions. *Journal of Personality and Social Psychology*, 63(6):932–944, 1992.

[98] Nilanjana Dasgupta, Melissa McManus Scircle, and Matthew Hunsinger. Female peers in small work groups enhance women's motivation, verbal participation, and career aspirations in engineering. *Proceedings of the National Academy of Sciences*, 112(16):4988–4993, 2015.

[99] Tara C Dennehy and Nilanjana Dasgupta. Female peer mentors early in college increase women's positive academic experiences and retention in engineering. *Proceedings of the National Academy of Sciences*, 114(23):5964–5969, 2017.

[100] Cissy J Ballen, Shima Salehi, and Sehoya Cotner. Exams disadvantage women in introductory biology. *PLOS ONE*, 12(10):1–14, 2017.

[101] Laura J Solomon and Esther D Rothblum. Academic procrastination: Frequency and cognitive-behavioral correlates. *Journal of Counseling Psychology*, 31(4):503–509, 1984.

[102] Guide to the 2018 ACT/SAT Concordance, 2018. https://www.act.org/content/dam/act/unsecured/documents/ACT-SAT-Concordance-Information.pdf.

[103] Marcus Credé and L Alison Phillips. A meta-analytic review of the motivated strategies for learning questionnaire. *Learning and Individual Differences*, 21(4):337–346, 2011.

[104] Paul R Pintrich, David A F Smith, Teresa Garcia, and Wilbert J Mckeachie. Reliability and predictive validity of the motivated strategies for learning questionnaire (MSLQ). *Educational and Psychological Measurement*, 53(3):801–813, 1993.

[105] Paul R Pintrich. *A Manual for the Use of the Motivated Strategies for Learning Questionnaire.* University of Michigan, 1991.

[106] Li-tze Hu and Peter M Bentler. Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological methods*, 3(4):424–453, 1998.

[107] Michael W Browne and Robert Cudeck. Alternative ways of assessing model fit. *Sage Focus Editions*, 154:136–136, 1993.

[108] Bruce B Frey. *The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation.* SAGE Publications, Inc., Thousand Oaks, CA, 2018.

[109] Jacob Cohen, Patricia Cohen, Stephen West, and Leona S Aiken. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences.* Lawrence Erlbaum, 2003.

[110] Dustin Tingley, Teppei Yamamoto, Kentaro Hirose, Luke Keele, and Kosuke Imai. Mediation: R package for causal mediation analysis. *Journal of Statistical Software*, 59(5):1–38, 2014.

[111] The College Board. *2020 Total Group SAT Suite of Assessments Annual Report.* The College Board, 2020. https://reports.collegeboard.org/media/pdf/2020-total-group-sat-suite-assessments-annual-report.pdf.

[112] Brian M Galla, Elizabeth P Shulman, Benjamin D Plummer, Margo Gardner, Stephen J Hutt, J Parker Goyer, Sidney K D'Mello, Amy S Finn, and Angela 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. *American Educational Research Journal*, 56(6):2077–2115, 2019.

[113] Zahra Hazari, Robert H Tai, and Philip M Sadler. Gender differences in introductory university physics performance: The influence of high school physics preparation and affective factors. *Science Education*, 91(6):847–876, 2007.

[114] Jayson M Nissen and Jonathan T Shemwell. Gender, experience, and self-efficacy in introductory physics. *Physical Review Physics Education Research*, 12(2):020105, 2016.

[115] Irene F Goodman. Final Report of the Women's Experiences in College Engineering (WECE) Project. 2002. https://eric.ed.gov/?id=ED507394.

[116] Sian L Beilock, Robert J Rydell, and Allen R McConnell. Stereotype threat and working memory: Mechanisms, alleviation, and spillover. *Journal of Experimental Psychology: General*, 136(2):256–276, 2007.

[117] Ronald E Smith. Effects of coping skills training on generalized self-efficacy and locus of control. *Journal of Personality and Social Psychology*, 56(2):228–233, 1989.

[118] Judith L Meece, Allan Wigfield, and Jacquelynne S Eccles. Predictors of math anxiety and its influence on young adolescents' course enrollment intentions and performance in mathematics. *Journal of Educational Psychology*, 82(1):60–70, 1990.

[119] James T Laverty, Wolfgang Bauer, Gerd Kortemeyer, and Gary Westfall. Want to reduce guessing and cheating while making students happier? Give more exams! *The Physics Teacher*, 50(9):540–543, 2012.

[120] David C Haak, Janneke Hillerislambers, Emile Pitre, and Scott Freeman. Increased structure and active learning reduce the achievement gap in introductory biology. *Science*, 332(6034):1213–1216, 2011.

[121] Jessica E Bickel, Leah M Bunnell, and Thijs Heus. Utilizing peer teaching and reflection on low-stakes quizzes to improve concept learning outcomes in introductory calculus-based physics classes. *European Journal of Physics*, 42(5):055701, 2021.

[122] John R Anderson. *Learning and Memory: An Integrated Approach 2nd Edition*. Wiley, 2000.

[123] David C Haak, Janneke Hillerislambers, Emile Pitre, and Scott Freeman. Increased structure and active learning reduce the achievement gap in introductory biology. *Science*, 332(6034):1213–1216, 2011.

[124] Scott Freeman, Sarah L Eddy, Miles McDonough, Michelle K Smith, Nnadozie Okoroafor, Hannah Jordt, and Mary Pat Wenderoth. Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences*, 111(23):8410–8415, 2014.

[125] Chandralekha Singh. Assessing student expertise in introductory physics with isomorphic problems. I. Performance on nonintuitive problem pair from introductory physics. *Physical Review Physics Education Research*, 4(1):010104, 2008.

[126] Chandralekha Singh. Assessing student expertise in introductory physics with isomorphic problems. II. Effect of some potential factors on problem solving and transfer. *Physical Review Physics Education Research*, 4(1):010105, 2008.

[127] Albert Bandura, W H Freeman, and Richard Lightsey. Self-efficacy: The exercise of control. *Journal of Cognitive Psychotherapy*, 13(2):158–166, 1999.

[128] Angela J Little, Bridget Humphrey, Abigail Green, Abhilash Nair, and Vashti Sawtelle. Exploring mindset's applicability to students' experiences with challenge in transformed college physics courses. *Physical Review Physics Education Research*, 15(1):010127, 2019.

[129] Tobias Espinosa, Kelly Miller, Ives Araujo, and Eric Mazur. Reducing the gender gap in students' physics self-efficacy in a team-and project-based introductory physics class. *Physical Review Physics Education Research*, 15(1):010132, 2019.

[130] Jianlan Wang and Zahra Hazari. Promoting high school students' physics identity through explicit and implicit recognition. *Physical Review Physics Education Research*, 14(2):020111, 2018.

[131] Zahra Hazari and Cheryl Cass. Towards meaningful physics recognition: What does this recognition actually look like? *The Physics Teacher*, 56(7):442–446, 2018.

[132] Kevin R Binning, Nancy Kaufmann, Erica M McGreevy, Omid Fotuhi, Susie Chen, Emily Marshman, Z Yasemin Kalender, Lisa Limeri, Laura Betancur, and Chandralekha Singh. Changing social contexts to foster equity in college science courses: An ecological-belonging intervention. *Psychological Science*, 31(9):1059–1070, 2020.

[133] Eric Brewe, Vashti Sawtelle, Laird H Kramer, George E O'Brien, Idaykis Rodriguez, and Priscilla Pamelá. Toward equity through participation in modeling instruction in introductory university physics. *Physical Review Physics Education Research*, 6(1):010106, 2010.

[134] Adrienne Traxler and Eric Brewe. Equity investigation of attitudinal shifts in introductory physics. *Physical Review Physics Education Research*, 11(2):020132, 2015.

[135] Ben Van Dusen and Jayson Nissen. Equity in college physics student learning: A critical quantitative intersectionality investigation. *Journal of Research in Science Teaching*, 57(1):33–57, 2020.

[136] Adrienne L Traxler, Ximena C Cid, Jennifer Blue, and Ramón Barthelemy. Enriching gender in physics education research: A binary past and a complex future. *Physical Review Physics Education Research*, 12:020114, 2016.

[137] Paul R Pintrich and Dale H Schunk. *Motivation in education: Theory, research, and applications*. Prentice Hall, 2002.

[138] Brian M Galla, Elizabeth P Shulman, Benjamin D Plummer, Margo Gardner, Stephen J Hutt, J. Parker Goyer, Sidney K D'Mello, Amy S Finn, and Angela 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. *American Educational Research Journal*, 56(6):2077–2115, 2019.

[139] Paulette Vincent-Ruz, Kevin Binning, Christian Schunn, and Joe Grabowski. The effect of math SAT on women's chemistry competency beliefs. *Chemistry Education Research and Practice*, 19:342–351, 2018.

[140] Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. L. Erlbaum Associates, Hillsdale, N.J., 1988.

[141] Jacob Cohen, Patricia Cohen, Stephen West, and Leona S Aiken. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. Lawrence Erlbaum, 2003.

[142] Irene Marzoli, Arturo Colantonio, Claudio Fazio, Marco Giliberti, Umberto Scotti di Uccio, and Italo Testa. Effects of emergency remote instruction during the COVID-19 pandemic on university physics students in Italy. *Physical Review Physics Education Research*, 17:020130, 2021.

[143] Elina Palmgren, Kimmo Tuominen, and Inkeri Kontro. Self-efficacy and conceptual knowledge in quantum mechanics during teaching reforms and the COVID-19 pandemic. *Physical Review Physics Education Research*, 18:020122, 2022.

[144] Gerd Kortemeyer, Wolfgang Bauer, and Wade Fisher. Hybrid teaching: A tale of two populations. *Physical Review Physics Education Research*, 18:020130, 2022.

[145] Stefan Klumpp, Sarah Köster, Anne C Pawsey, Yvonne Lips, Martin Wenderoth, and Pascal Klein. Reflections on COVID-19–Induced Online Teaching in Biophysics Courses. *The Biophysicist*, 2(2):20–22, 2021.

[146] Ben Van Dusen, Mollee Shultz, Jayson M Nissen, Bethany R Wilcox, N. G Holmes, Manher Jariwala, Eleanor W Close, H. J Lewandowski, and Steven Pollock. Online administration of research-based assessments. *American Journal of Physics*, 89(1):7–8, 2021.

[147] Zhongzhou Chen. Measuring the level of homework answer copying during covid-19 induced remote instruction. *Physical Review Physics Education Research*, 18:010126, 2022.

[148] Chandralekha Singh. Impact of peer interaction on conceptual test performance. *American Journal of Physics*, 73(5):446–451, 2005.

[149] Peter Hu, Yangqiuting Li, and Chandralekha Singh. Challenges in addressing student difficulties with measurement uncertainty of two-state quantum systems using a multiple-choice question sequence in online and in-person classes. *European Journal of Physics*, 44:015702, 2023.

[150] Elise Swanson, Tatiana Melguizo, and Paco Martorell. Examining the relationship between psychosocial and academic outcomes in higher education: A descriptive analysis. *AERA Open*, 7:23328584211026967, 2021.

[151] Jessi L Smith, Karyn L Lewis, Lauren Hawthorne, and Sara D Hodges. When trying hard isn't natural: Women's belonging with and motivation for male-dominated STEM fields as a function of effort expenditure concerns. *Personality and Social Psychology Bulletin*, 39(2):131–143, 2013.

[152] Rebecca L Matz, Benjamin P Koester, Stefano Fiorini, Galina Grom, Linda Shepard, Charles G Stangor, Brad Weiner, and Timothy A McKay. Patterns of gendered performance differences in large introductory courses at five research universities. *AERA Open*, 3(4):2332858417743754, 2017.

[153] Gaydaa Al-Zohbi, Maura A E Pilotti, Kamal Barghout, Omar Elmoussa, and Hanadi Abdelsalam. Lesson learned from the pandemic for learning physics. *Journal of Computer Assisted Learning*, 2022.

[154] StataCorp. *Stata Statistical Software: Release 17.* StataCorp LLC., College Station, TX, 2021.

[155] David MacKinnon. *Introduction to Statistical Mediation Analysis.* Taylor & Francis Group, Mahwah, NJ, 2008.

[156] Barbara King. Does postsecondary persistence in STEM vary by gender? *AERA Open*, 2(4):2332858416669709, 2016.

[157] Stepfanie M Aguillon, Gregor-Fausto Siegmund, Renee H Petipas, Abby Grace Drake, Sehoya Cotner, and Cissy J Ballen. Gender differences in student participation in an active-learning classroom. *CBE—Life Sciences Education*, 19(2):ar12, 2020.

[158] Ben Van Dusen and Jayson Nissen. Equity in college physics student learning: A critical quantitative intersectionality investigation. *Journal of Research in Science Teaching*, 57(1):33–57, 2020.

[159] Jessica H Dwyer, Wilson J González-Espada, Kimberly de la Harpe, and David C Meier. Factors associated with students graduating with STEM de-

grees at a military academy: Improving success by identifying early obstacles. *Journal of College Science Teaching*, 50(1):28–35, 2020.

[160] Zahra Hazari, Phillip M Sadler, and Gerhard Sonnert. The science identity of college students: Exploring the intersection of gender, race, and ethnicity. *Journal of College Science Teaching*, 42(5):82–91, 2013.

[161] Jennifer McKinney, Mei-Lin Chang, and David Glassmeyer. Why females choose STEM majors: Understanding the relationships between major, personality, interests, self-efficacy, and anxiety. *Journal for STEM Education Research*, 4(3):278–300, 2021.

[162] Nilanjana Dasgupta. Ingroup experts and peers as social vaccines who inoculate the self-concept: The stereotype inoculation model. *Psychological Inquiry*, 22(4):231–246, 2011.

[163] Jane G Stout, Tiffany A Ito, Noah D Finkelstein, and Steven J Pollock. How a gender gap in belonging contributes to the gender gap in physics participation. In *AIP Conference Proceedings*, volume 1513, pages 402–405, Melville, New York. American Institute of Physics.

[164] Benson Adesina Adegoke. Impact of interactive engagement on reducing the gender gap in quantum physics learning outcomes among senior secondary school students. *Physics Education*, 47(4):462–470, 2012.

[165] Chandralekha Singh. Effectiveness of group interaction on conceptual standardized test performance. In *Physics Education Research Conference 2002*, PER Conference, Biose, Idaho, August 7-8 2002. https://doi.org/10.1119/perc.2002.pr.017.

[166] Packard Wai-Ling Becky, Jaemarie Solyst, Anisha Pai, and Lu Yu. Peer-designed active learning modules as a strategy to improve confidence and comprehension within introductory computer science. *Journal of College Science Teaching*, 49(5):76–83, 2020.

[167] Fosdick Bailey K Rasmussen Chris Ellis, Jessica. Women 1.5 times more likely to leave stem pipeline after calculus compared to men: Lack of mathematical confidence a potential culprit. *PloS one*, 11(7):e0257447, 2020.

[168] Triaka Larry and Jillian L Wendt. Predictive relationship between gender, ethnicity, science self-efficacy, teacher interpersonal behaviors, and science achievement of students in a diverse urban high school. *Learning Environments Research*, 25:141–157, 2021.

[169] Chandralekha Singh. Prior preparation and motivational characteristics mediate relations between gender and learning outcomes in introductory physics. In *Physics Education Research Conference 2018*, PER Conference, Washington, DC., August 1-2 2018. https://doi.org/10.1119/perc.2018.pr.Nokes-Malach.

[170] Patricia Heller and Mark Hollabaugh. Teaching problem solving through cooperative grouping. Part 2: Designing problems and structuring groups. *American Journal of Physics*, 60(7):637–644, 1992.

[171] Melanie Good, Alexandru Maries, and Chandralekha Singh. Impact of traditional or evidence-based active-engagement instruction on introductory female and male students' attitudes and approaches to physics problem solving. *Physical Review Physics Education Research*, 15(2):020129, 2019.

[172] L Hu and P M Bentler. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1):1–55, 1999.

[173] Rex B Kline. *Principles and Practice of Structural Equation Modeling.* Guilford Press, New York, 4th edition, 2016.

[174] R Core Team. R: A language and environment for statistical computing, 2020.

[175] Yves Rosseel. lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2):1–36, 2012.

[176] Denise Wilson, Diane Jones, Fraser Bocell, Joy Crawford, Mee Joo Kim, Nanette Veilleux, Tamara Floyd-Smith, Rebecca Bates, and Melani Plett. Belonging and academic engagement among undergraduate STEM students: A multi-institutional study. *Research in Higher Education*, 56(7):750–776, 2015.

[177] Anne Deiglmayr, Elsbeth Stern, and Renate Schubert. Beliefs in "brilliance" and belonging uncertainty in male and female stem students. *Frontiers in Psychology*, 10:1114, 2019.

[178] Michael Briggs. Comparing academically homogeneous and heterogeneous groups in an active learning physics class. *Journal of College Science Teaching*, 49(6):76–82, 2020.

[179] Michael Brown and Robert M DeMonbrun. Who gets helped? the opportunity structure of the college physics classroom, peer instruction, and perceptions of help seeking. *Journal of College Science Teaching*, 41(2):36–44, 2019.

[180] Alexandru Maries, Nafis I Karim, and Chandralekha Singh. Active learning in an inequitable learning environment can increase the gender performance gap: The negative impact of stereotype threat. *The Physics Teacher*, 58:430–433, 2020.

[181] Kevin Crowley, Maureen A Callanan, Harriet R Tenenbaum, and Elizabeth Allen. Parents explain more often to boys than to girls during shared scientific thinking. *Psychological Science*, 12(3):258–261, 2001.

[182] Daniel Z Grunspan, Sarah L Eddy, Sara E Brownell, Benjamin L Wiggins, Alison J Crowe, and Steven M Goodreau. Males under-estimate academic performance of their female peers in undergraduate biology classrooms. *PLoS ONE*, 11:e0148405, 2016.

[183] Allison J Gonsalves, Anna Danielsson, and Helena Pettersson. Masculinities and experimental practices in physics: The view from three case studies. *Physical Review Physics Education Research*, 12(2):020120, 2016.

[184] Frank Pajares. Current directions in self-efficacy research. *Advances in Motivation and Achievement*, 10:1–49, 1997.

[185] Zahra Hazari, Gerhard Sonnert, Philip M Sadler, and Marie-Claire Shanahan. Connecting high school physics experiences, outcome expectations, physics identity, and physics career choice: A gender study. *Journal of Research in Science Teaching*, 47(8), 2010.

[186] Melvin Vooren, Carla Haelermans, Wim Groot, and Henriette Maassen van den Brink. Comparing success of female students to their male counterparts in the STEM fields: An empirical analysis from enrollment until graduation using longitudinal register data. *International Journal of STEM Education*, 9(1), 2022. https://doi.org/10.1186/s40594-021-00318-8.

[187] Lauren E Kost, Steven J Pollock, and Noah D. Finkelstein. Characterizing the gender gap in introductory physics. *Physical Review Physics Education Research*, 5:010101, 2009.

[188] Joseph A Raelin, Margaret B Bailey, Jerry Hamann, Leslie K Pendleton, Rachelle Reisberg, and David L Whitman. The gendered effect of cooperative education, contextual support, and self-efficacy on undergraduate retention. *Journal of Engineering Education*, 103(4):599–624, 2014.

[189] Rachel E Scherr, Monica Plisch, Kara E Gray, Geoff Potvin, and Theodore Hodapp. Fixed and growth mindsets in physics graduate admissions. *Physical Review Physics Education Research*, 13:020133, 2017.

[190] Kelly Miller, Julie Schell, Andrew Ho, Brian Lukoff, and Eric Mazur. Response switching and self-efficacy in peer instruction classrooms. *Physical Review Physics Education Research*, 11:010104, 2015.

[191] D S Yeager and C S Dweck. What can be learned from growth mindset controversies? *American Psychologist*, 75(9):1269–1284, 2020.

[192] David A Cook, Richmond M Castillo, Becca Gas, and Anthony R. Artino Jr. Measuring achievement goal motivation, mindsets and cognitive load: validation of three instruments' scores. *Medical Education*, 51(10):1061–1074, 2017. https://doi.org/10.1186/s40594-020-00219-2.

[193] Shu-Shen Shih. Perfectionism, implicit theories of intelligence, and taiwanese eighth-grade students' academic engagement. *The Journal of Educational Research*, 104(2):131–142, 2011.

[194] Stefan J Troche and Alexandra Kunz. The factorial structure and construct validity of a german translation of dweck's implicit theories of intelligence scale under consideration of the wording effect. *Psychology Science*, 62(3):386–403, 2020.

[195] Katherine Kricorian, Michelle Seu, Daniel Lopez, Elsie Ureta, and Ozlem Equils. Factors influencing participation of underrepresented students in stem fields: matched mentors and mindsets. *International Journal of STEM Education*, 7(1):16, 2020. https://doi.org/10.1186/s40594-020-00219-2.

[196] Céline Bagès, Catherine Verniers, and Delphine Martinot. Virtues of a hard-working role model to improve girls' mathematics performance. *Psychology of Women Quarterly*, 40(1):55–64, 2015.

[197] D S Yeager, C Romero, D Paunesku, C S Hulleman, B Schneider, C Hinojosa, H Y Lee, J O'Brien, K Flint, Roberts A, J Trott, D Greene, G M Walton, and C S Dweck. Using design thinking to improve psychological interventions: The case of the growth mindset during the transition to high school. *Journal of Educational Psychology*, 108(3):374–391, 2016.

[198] David S Yeager, Paul Hanselman, Gregory M Walton, Jared S Murray, Robert Crosnoe, Chandra Muller, Elizabeth Tipton, Barbara Schneider, Chris S Hulleman, Cintia P Hinojosa, David Paunesku, Carissa Romero, Kate Flint, Alice Roberts, Jill Trott, Ronaldo Iachan, Jenny Buontempo, Sophia Man Yang, Carlos M Carvalho, P Richard Hahn, Maithreyi Gopalan, Pratik Mhatre, Ronald Ferguson, Angela L Duckworth, and Carol S Dweck. A national

experiment reveals where a growth mindset improves achievement. *Nature*, 573(7774):364–369, 2019.

[199] Victoria F Sisk, Alexander P Burgoyne, Jingze Sun, Jennifer L Butler, and Brooke N Macnamara. To what extent and under which circumstances are growth mindsets important to academic achievement? Two meta-analyses. *Psychological Science*, 29(4):549–571, 2018.

[200] John Hattie. *Visible learning for teachers: Maximizing impact on learning.* Routledge, New York, New York, 2012.

[201] Krista De Castella and Donald Byrne. My intelligence may be more malleable than yours: The revised implicit theories of intelligence (self-theory) scale is a better predictor of achievement, motivation, and student disengagement. *European Journal of Psychology of Education*, 30(3):245–267, 2015.

[202] Colleen M Ganley, Casey E George, Joseph R Cimpian, and Martha B Makowski. Gender equity in college majors: Looking beyond the STEM/non-STEM dichotomy for answers regarding female participation. *American Educational Research journal*, 55(3):453–487, 2018.

[203] Daniel Voyer and Susan D. Voyer. Gender differences in scholastic achievement: A meta-analysis. *Psychological Bulletin*, 140(4):1174–1204, 2014.

[204] Caitlin Kepple, Marakee Tilahun, Natalia Matti, and Kim Coble. Pedagogy training for the development of GTA mindsets and inclusive teaching practices. In *Physics Education Research Conference 2020*, PER Conference, pages 272–277, Virtual Conference, July 22-23 2020.

[205] Zach C Schudson. Psychology's stewardship of gender/sex. *Perspectives on Psychological Science*, 16(6):1105–1112, 2021.

[206] Paulette Vincent-Ruz and Christian D Schunn. The nature of science identity and its role as the driver of student choices. *International Journal of STEM Education*, 5(1):1–12, 2018. ttps://doi.org/10.1186/s40594-018-0140-5.

[207] María Garcá-Cepero and D Betsy Mccoach. Educators' implicit theories of intelligence and beliefs about the identification of gifted students. *Universitas Psychologica*, 8:295–310, 2009.

[208] Corinne A Moss-Racusin, John F Dovidio, Victoria L Brescoll, Mark J Graham, and Jo Handelsman. Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences*, 109(41):16474, 2012.

[209] Claudia M Mueller and Carol S Dweck. Praise for intelligence can undermine children's motivation and performance. *Journal of personality and social psychology*, 75(1):33, 1998.

[210] Gregory M Walton and Geoffrey L Cohen. A question of belonging: Race, social fit, and achievement. *Journal of Personality and Social Psychology*, 92(1):82, 2007.

[211] Elizabeth A Canning, Katherine Muenks, Dorainne J Green, and Mary C Murphy. Stem faculty who believe ability is fixed have larger racial achievement gaps and inspire less student motivation in their classes. *Science Advances*, 5(2):eaau4734, 2019.

[212] Aneeta Rattan, Catherine Good, and Carol S Dweck. "It's OK—Not everyone can be good at math": Instructors with an entity theory comfort (and demotivate) students. *Journal of Experimental Social Psychology*, 48(3):731–737, 2012.

[213] Brandon J Yik, Jeffrey R Raker, Naneh Apkarian, Marilyne Stains, Charles Henderson, Melissa H Dancy, and Estrella Johnson. Evaluating the impact of malleable factors on percent time lecturing in gateway chemistry, mathemat-

ics, and physics courses. *International Journal of STEM Education*, 9(1):15, 2022. https://doi.org/10.1186/s40594-022-00333-3.

[214] J LaCosse, M C Murphy, J A Garcia, and S Zirkel. The role of stem professors' mindset beliefs on students' anticipated psychological experiences and course interest. *Journal of Educational Psychology*, 113(5):949–971, 2021.

[215] Richard M Felder, Gary N Felder, Meredith Mauney, Charles E Hamrin Jr., and E Jacquelin Dietz. A longitudinal study of engineering student performance and retention. III. Gender differences in student performance and attitudes. *Journal of Engineering Education*, 84(2):151–163, 1995.

[216] Elaine Howard Ecklund, Anne E Lincoln, and Cassandra Tansey. Gender segregation in elite academic science. *Gender and Society*, 26(5):693–717, 2012.

[217] Luke Holman, Devi Stuart-Fox, and Cindy E Hauser. The gender gap in science: How long until women are equally represented? *PLOS Biology*, 16(4):1–20, 2018.

[218] Angela Johnson, Maria Ong, Lily T Ko, Janet Smith, and Apriel Hodari. Common challenges faced by women of color in physics, and actions faculty can take to minimize those challenges. *The Physics Teacher*, 55(6):356–360, 2017.

[219] Advanced Placement 2021 program summary report, 2021. https://reports.collegeboard.org/media/pdf/2021-ap-program-summary-report_1.pdf.

[220] Sapna Cheryan and Victoria C Plaut. Explaining underrepresentation: A theory of precluded interest. *Sex Roles*, 63:475–488, 2010.

[221] David Scott Yeager, Valerie Purdie-Vaughns, Julio Garcia, Nancy Apfel, Patti Brzustoski, Allison Master, William T Hessert, Matthew E Williams, and

Geoffrey L Cohen. Breaking the cycle of mistrust: Wise interventions to provide critical feedback across the racial divide. *Journal of Experimental Psychology: General*, 143(2):804, 2014.

[222] Sandra Upson and Lauren F Friedman. Where are all the female geniuses? *SA Mind*, 23:63–65, 2012.

[223] R C MacCallum and H M Browne, M W ans Sugawara. Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, 1(2):130–149, 1996.

[224] Zahra Hazari, Deepa Chari, Geoff Potvin, and Eric Brewe. The context dependence of physics identity: Examining the role of performance/competence, recognition, interest, and sense of belonging for lower and upper female physics undergraduates. *Journal of Research in Science Teaching*, 57(10):1583–1607, 2020.

[225] Ramón S Barthelemy, Melinda McCormick, and Charles Henderson. Gender discrimination in physics and astronomy: Graduate student experiences of sexism and gender microaggressions. *Physical Review Physics Education Research*, 12:020119, 2016.

[226] Katemari Rosa and Felicia Moore Mensah. Educational pathways of black women physicists: Stories of experiencing and overcoming obstacles in life. *Physical Review Physics Education Research*, 12:020113, 2016.

[227] Julie R Posselt, Theresa E Hernandez, Geraldine L Cochran, and Casey W Miller. Metrics first, diversity later? Making the short list and getting admitted to physics PhD programs. *Journal of Women and Minorities in Science and Engineering*, 25(4):283–306, 2019.

[228] Geraldine L Cochran, Theodore Hodapp, and Erika E Alexander Brown. Identifying barriers to ethnic/racial minority students' participation in graduate physics. In *Physics Education Research Conference 2017*, PER Conference, pages 92–95, Cincinnati, OH, July 26-27 2017.

[229] Westley James, Caroline Bustamante, Kamryn Lamons, Erin Scanlon, and Jacquelyn J. Chini. Disabling barriers experienced by students with disabilities in postsecondary introductory physics. *Physical Review Physics Education Research*, 16:020111, 2020.

[230] Amanda Lannan, Jacquelyn J Chini, and Erin Scanlon. Resources for supporting students with and without disabilities in your physics courses. *The Physics Teacher*, 59(3):192–195, 2021.

[231] Becky Francis, Louise Archer, Julie Moote, Jen DeWitt, Emily MacLeod, and Lucy Yeomans. The construction of physics as a quintessentially masculine subject: Young people's perceptions of gender issues in access to physics. *Sex Roles*, 76:156–174, 2016.

[232] Michela Musto. Brilliant or bad: The gendered social construction of exceptionalism in early adolescence. *American Sociological Review*, 84:369–393, 2019.

[233] Burkhard Gniewosz, Jacquelynne S Eccles, and Peter Noack. Early adolescents' development of academic self-concept and intrinsic task value: The role of contextual feedback. *Journal of Research on Adolescence*, 25(3):459–473, 2015.

[234] Kevin Rask. Attrition in STEM fields at a liberal arts college: The importance of grades and pre-collegiate preferences. *Economics of Education Review*, 29(6):892–900, 2010.

[235] Benjamin P Koester, Galina Grom, and Timothy A. McKay. Patterns of gendered performance difference in introductory STEM courses, 2016. https://doi.org/10.48550/arXiv.1608.07565.

[236] Allan Wigfield and Jenna Cambria. Students' achievement values, goal orientations, and interest: Definitions, development, and relations to achievement outcomes. *Developmental Review*, 30(1):1–35, 2010.

[237] Kevin Rask and Jill Tiefenthaler. The role of grade sensitivity in explaining the gender imbalance in undergraduate economics. *Economics of Education Review*, 27(6):676–687, 2008.

[238] Madeline Huberth, Patricia Chen, Jared Tritz, and Timothy A. McKay. Computer-tailored student support in introductory physics. *PLOS ONE*, 10(9):e0137001, 2015.

[239] Dirk Witteveen and Paul Attewell. The STEM grading penalty: An alternative to the "leaky pipeline" hypothesis. *Science Education*, 104(4):714–735, 2020.

[240] Cissy J Ballen, Carl Wieman, Shima Salehi, Jeremy B Searle, and Kelly R. Zamudio. Enhancing diversity in undergraduate science: Self-efficacy drives performance gains with active learning. *CBE—Life Sciences Education*, 16(4):ar56, 2017.

[241] Amy L Zeldin, Shari L Britner, and Frank Pajares. A comparative study of the self-efficacy beliefs of successful men and women in mathematics, science, and technology careers. *Journal of Research in Science Teaching*, 45(9):1036–1058, 2008.

[242] Rachel A Louis and Jean M Mistele. The differences in scores and self-efficacy by student gender in mathematics and science. *International Journal of Science and Mathematics Education*, 10(5):1163–1190, 2012.

[243] Shelley J. Correll. Gender and the career choice process: The role of biased self-assessments. *American Journal of Sociology*, 106:1691–1730, 2001.

[244] Shelley J. Correll. Constraints into preferences: Gender, status, and emerging career aspirations. *American Sociological Review*, 69:93–113, 2004.

[245] Jim Lemon. Plotrix: a package in the red light district of R. *R-News*, 6(4):12, 2006.

[246] Danielle Navarro. Learning statistics with R: A tutorial for psychology students and other beginners, 2015.

[247] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag, New York, 2016.

[248] Marissa E. Thompson. Grade expectations: The role of first-year grades in predicting the pursuit of STEM majors for first- and continuing-generation students. *The Journal of Higher Education*, 92(6):961–985, 2021.

[249] Nafis I Karim, Alexandru Maries, and Chandralekha Singh. Impact of evidence-based flipped or active-engagement non-flipped courses on student performance in introductory physics. *Canadian Journal of Physics*, 96(4):411–419, 2018.

[250] Catherine Good, Aneeta Rattan, and Carol S Dweck. Why do women opt out? Sense of belonging and women's representation in mathematics. *Journal of Personality and Social Psychology*, 102(4):700, 2012.

[251] Sarah L Eddy, Sara E Brownell, and Mary Pat Wenderoth. Gender gaps in achievement and participation in multiple introductory biology classrooms. *CBE Life Sciences Education*, 13:478–492, 2014.

[252] 2021 FACTS: Applicants and Matriculants Data. Report, 2021.

[253] Samantha L Elliott. From the editor-in-chief: Questions of gender equity in the undergraduate biology classroom. *Journal of Microbiology Biology Education*, 17:186–188, 2016.

[254] Valerie M Dandar and Diana M Lautenberger. Exploring Faculty Salary Equity at U.S. Medical Schools by Gender and Race/Ethnicity. Report, 2021. https://www.aamc.org/data-reports/faculty-salary-equity.

[255] Dan P Ly, Seth A Seabury, Anupam B Jena, and Ruth L Newhouse. Differences in incomes of physicians in the united states by race and sex: Observational study. *BMJ*, 353:i2923, 2016.

[256] Junming Huang, Alexander J Gates, Roberta Sinatra, and Albert-Lászlo Barabási. Historical comparison of gender inequality in scientific careers across countries and disciplines. *Proceedings of the National Academy of Sciences*, 117(9):4609–4616, 2020.

[257] Martin Husemann, Rebecca Rogers, Sebastian Meyer, and Jan Christian Habel. "publicationism" and scientists' satisfaction depend on gender, career stage and the wider academic system. *Palgrave Communications*, 3:17032, 2017.

[258] Roy Y Chan, Krishna Bista, and Ryan M Allen, editors. *Online Teaching and Learning in Higher Education During COVID-19 : International Perspectives and Experiences.* Routledge, New York, NY, 2022.

[259] Bob Ives and Ana-Maria Cazan. Did the COVID-19 pandemic lead to an increase in academic misconduct in higher education? (online ahead of print). *Higher Education*, pages 1–19, 2023. https://doi.org/10.1007/s10734-023-00996-z.

[260] T Gonzalez, M A de la Rubia, K.P. Hincz, M Comas-Lopez, Laia Subirats, Santi Fort, and G M Sacha. Influence of covid-19 confinement on students' performance in higher education. *PLOS ONE*, 15(10):e0239490., 2020.