Investigating Gender Differences in Course Relationships, Self-Efficacy, and Identity in Physics and Promoting Equity in Learning Outcomes

by

Kyle Maclane Whitcomb

B.S., Mathematics, University of Puget Sound, 2015

B.S., Physics, University of Puget Sound, 2015

Submitted to the Graduate Faculty of

the Kenneth P. Dietrich School of Arts and Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2020

UNIVERSITY OF PITTSBURGH

KENNETH P. DIETRICH SCHOOL OF ARTS AND SCIENCES

This dissertation was presented

by

Kyle Maclane Whitcomb

It was defended on

July 14th, 2020

and approved by

Dr. Chandralekha Singh, Professor, Department of Physics and Astronomy

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

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

Dr. Arthur Kosowsky, Professor, Department of Physics and Astronomy

Dr. Larry Shuman, Professor, Department of Industrial Engineering

Investigating Gender Differences in Course Relationships, Self-Efficacy, and Identity in Physics and Promoting Equity in Learning Outcomes

Kyle Maclane Whitcomb, PhD

University of Pittsburgh, 2020

Research in physics education faces a challenge of low sample size when focusing on issues pertaining to physics majors. However, with the continued collection of data over the past decades we are now reaching a point where there is sufficient statistical power to investigate crucial issues such as inequities within the physics curriculum. In recent years there has been a push towards identifying these inequities, especially with regards to gender or race, in introductory physics classes. The research presented here extends this work by utilizing 15 years of institutional data and 5 years of survey data in order to reveal inequities in the education of physics majors. These investigations primarily consider gender differences wherein we find troubling trends in introductory physics grades that may be a factor in women's decisions whether or not to pursue a major in physics. In addition, parallel investigations of engineering students provide context of how these same gender differences in introductory physics are entirely opposed to the trends of non-physics majors in other STEM disciplines and may be driven by stereotype-driven gender differences in physics self-efficacy. We further find in an investigation of the motivational characteristics of physics majors that there is a decline in physics majors' self-efficacy and physics identity over time, both during the tumultuous first year of study and even as the physics students progress to their final year of study. Moreover, investigations of STEM enrollment patterns will further situate physics as having a uniquely inequitable environment for women and underrepresented minority students, where student attrition from physics is more prevalent than in any other STEM major. Finally, we present a study of self-paced online learning tools that can be useful in future investigations of how these inequities influence learning throughout a course, which can provide even more direct and immediate feedback to instructors.

Table of Contents

List of Tables

List of Figures

Preface

I am eternally thankful to my parents, who inspired me to challenge myself and pursue a higher education. I am eternally grateful to my fiancée, Theresa, for her sacrifices that made this pursuit possible and for being an unwavering foundation for me through my graduate education.

I thank my Committee members for their invaluable discussions and feedback that helped to shape and focus my research.

I thank the National Science Foundation for their continued financial support of my research via Grant DUE-1524575.

I have many collaborators to thank, whose guidance and mentorship were critical foundations of my work. I thank Dr. Yasemin Kalender for her mentorship while I was first becoming acquainted with physics education research. I also thank Dr. Christian Schunn for introducing me to many statistical techniques that became the backbone of my research. Additionally, I thank Dr. Timothy Nokes-Malach for imparting his valuable knowledge of student learning. I am extremely grateful to Dr. Zhongzhou Chen for the great amount of time he spent mentoring me on physics education research and statistical analysis as well as using the R programming language for analysis and visualizations, which greatly increased the quality and speed of my research.

Last, but not least by any measure, I would like to profusely thank my research advisor, Dr. Chandralekha Singh. None of my research would be possible without Dr. Singh's expertise, guidance, and support. Her tireless work and passion for equity in physics education made my research the best it could be.

1.0 Introduction

1.1 Frameworks that Inspired this Research

1.1.1 Physics Identity and Its Components

The identity of a physics major as a “physics person” plays a major role in their academic trajectory and future career choices [73, 102, 142, 143, 114, 164]. Hazari and colleagues [115, 116] developed a framework of how physics identity depends on three motivational characteristics:

- perceived recognition by instructors, TAs, peers, and family (or “external identity”),
- physics self-efficacy (to be defined in the following paragraph), and
- interest in physics.

Prior studies have found that perceived recognition is the strongest predictor of physics identity, which implies that physics instructors can boost their students’ identification as a physics person by making an effort to recognize students as people who can excel in physics [115, 116, 142, 143, 145].

Self-efficacy, defined by Bandura as one’s belief in one’s ability to succeed at a particular task or subject [8, 9, 10, 11, 12, 13], is another extremely important motivational construct in this framework. Many prior studies of self-efficacy have shown that it is linked to many aspects of education, both in general and in physics specifically, including academic achievement, persistence in the major, and career choices after graduation [8, 9, 10, 11, 12, 13, 18, 24, 46, 144, 145, 157, 176, 177, 178, 214, 253, 254, 272, 292]. Low self-efficacy can be influenced by societal stereotypes and biases that produce stress and anxiety for students, and especially for women in physics, via “stereotype threat” (i.e., the fear of confirming stereotypes) which is not experienced by their male peers [4, 31, 46, 63, 95, 94, 126, 136]. In particular, stereotypes and biases about physics being a field for “brilliant men” cause stereotype threats for women that can increase anxiety in learning and test-taking situations and lead to deteriorated performance. Since anxiety can increase when performance deteriorates, these factors working against women in physics can force them into a feedback loop and hinder their performance further, which can further lower their self-efficacy and can continue to affect future performance [8, 9, 10, 11, 12, 13].

1.1.2 Expectancy-Value Theory

Expectancy value theory (EVT) states that a student's persistence and engagement in a discipline are related to student's expectancy about their success as well as how the student values the task [86, 87, 88]. In an academic context, "expectancy," which refers to the individual's beliefs about their success in the discipline, is closely related to Bandura's construct of self-efficacy [8, 9, 10, 11, 12, 13, 86, 87, 88].

There are four main factors that influence students' expectancy or self-efficacy, namely vicarious experiences (e.g., instructors or peers as role models), social persuasion (e.g., explicit mentoring, guidance, and support), level of anxiety [8, 9, 10, 11, 12, 13], and performance feedback (e.g., via grades on assessment tasks). Women generally have lower self-efficacy than men in many STEM disciplines, and especially in physics, because these four factors negatively influence them [14, 93, 136, 137, 213, 242, 261]. For example, women are underrepresented in calculus-based physics classrooms, and less likely to have a female role model among their physics instructors [199, 200]. Further, the stereotypes surrounding women in physics can affect how they are treated by mentors, even if such an effect is subconscious [4, 94, 95, 136, 63, 31, 46, 126]. Moreover, women are susceptible to stress and anxiety from stereotype threat which is not experienced by their male peers [4, 31, 46, 63, 94, 95, 126, 136]. This stress and anxiety can rob them of their cognitive resources, especially during high-stakes assessments such as exams.

In EVT, value is typically defined as having four facets: intrinsic value (i.e., interest in the task), attainment value (i.e., the importance of the task for the student's identity), utility value (i.e., the value of the task for future goals such as career), and cost (i.e., opportunity cost or psychological effects such as stress and anxiety) [86, 87, 88]. Intrinsic value can be informed by societal stereotypes and brilliance-attributions of physics, and attainment and utility values can be further tempered by these stereotypes. Utility value is an important facet of student education in STEM, since a degree in physics provides many job opportunities for graduating students. In addition, the psychological cost of majoring in these disciplines can be inflated by the stereotype threat. All of these effects can conspire to suppress the likelihood of women choosing and/or persisting in physics.

1.2 Physics Majors Throughout the Curriculum

In Chapter 2, we conduct an analysis of institutional data for physics majors showing predictive relationships between required mathematics and physics courses in various years, which are important for contemplating how the courses build on each other and whether there is need to make changes to the curriculum for the majors to strengthen these relationships. We used 15 years of institutional data at a US-based large research university to investigate how introductory physics and mathematics courses predict male and female physics majors' performance on required advanced physics and mathematics courses. We used Structure Equation Modeling (SEM) to investigate these predictive relationships and find that among introductory and advanced physics and mathematics courses, there are gender differences in performance in favor of male students only in the introductory physics courses after controlling for high school GPA. We found that a measurement invariance fully holds in a multi-group SEM by gender, so it was possible to carry out analysis with gender mediated by introductory physics and high school GPA. Moreover, we find that these introductory physics courses that have gender differences do not predict performance in advanced physics courses. In other words, students could be using invalid data about their introductory physics performance to make their decision about whether physics is the right field for them to pursue, and those invalid data in introductory physics favor male students. Also, introductory mathematics courses predict performance in advanced mathematics courses which in turn predict performance in advanced physics courses. Furthermore, apart from the introductory physics courses that do not predict performance in future physics courses, there is a strong predictive relationship between the sophomore, junior and senior level physics courses.

Chapter 3 explores the dependence of physics identity on other motivational characteristics for physics majors throughout a physics curriculum. The importance of motivational characteristics in physics education has been increasingly recognized in recent years, with prior studies establishing a framework for how students' identity as a physics person is composed of their perceived recognition by instructors and peers, along with their physics self-efficacy and interest. We seek to extend this research beyond introductory physics students by using five years of cross-sectional data collected from motivational surveys administered to physics majors throughout their undergraduate education and to the first year physics Ph.D. students at a large research university in the US. We find that physics majors in the first year tend to respond to the survey prompts more positively than their non-physics major peers, though with a smaller mean difference in self-efficacy than in the other motivational characteristics. Further, the responses of physics majors over time from their first year of the undergraduate curriculum through the first year of graduate school remain largely consistent, indicating that students are constantly adjusting their interpretation of the survey items to match the current level of expertise expected of them. Finally, we find that, consistent with prior studies with introductory physics students, perceived recognition is the best predictor of physics identity for physics majors throughout their entire education in physics, pointing to the importance of physics instructors making a concerted effort to constantly recognize their students throughout their studies.

1.3 The Experiences of Engineering Students in Physics

Beginning in Chapter 4, we explore how engineering students' experiences with physics relate to their experiences in other STEM courses. Math and science courses (physics, chemistry and mathematics) are considered foundational in engineering curricula and all engineering undergraduates must successfully complete courses in these subjects. However, relatively little is known about the predictive relationships between foundational math/science/engineering coursework and later engineering courses. This study uses large-scale institutional data to investigate the relationships between grades earned in foundational courses and early engineering courses in two large majors in order to gain insight into which foundational courses are most predictive of later performance and whether the relationship follows a linear or threshold function. Multiple regression analyses were performed on course grades using 10 years of data on 5,348 engineering students to construct a predictive model. We find that the predictive relationship between early and later performance is generally linear rather than threshold and that the strongest predictors are advanced mathematics courses along with cumulative STEM GPA, which is in turn strongly predicted by high school GPA and entry test scores. Physics and introductory engineering programming and modeling courses from the first year also predict performance in later courses. Advanced mathematics courses are critical to the long-term success of engineering students in these two common majors and students should be encouraged to aim for high rather than minimally passing grades.

Chapter 5 extends the research in Chapter 4 by considering how the grades earned by men and women in engineering programs differ in each of the STEM disciplines foundational to the engineering curriculum. We investigate gender differences in the predictive relationship between high school GPA and foundational mathematics and science courses as well as required engineering courses in the first two years for the undergraduate mechanical engineering majors using ten years of institutional data at a large research university. We use Structural Equation Modeling (SEM) to analyze the strength of these predictive relationships in order to understand the extent to which the curriculum is cohesive in the first two years and courses build on each other. We find a strong predictive pathway from high school GPA to overall first-year performance and then to the performance in advanced mathematics courses and finally to second-year mechanical engineering courses. Further, we use multi-group SEM in order to test for gender differences in individual courses as well as the predictive relationships between courses. We find that although high school GPA predicts first year course performance overall, women, who have a higher average high school GPA than men, underperform compared to men in first-year physics. First-year physics also displays the only difference in a predictive relationship, with overall first-year performance relating more strongly to physics performance for men than for women. These findings can be useful in evaluating the relationships between courses in the engineering curriculum. Further, understanding where gender differences occur in a curriculum is an important step towards promoting equity and inclusion and improving the learning environment so that all students can excel in engineering programs.

Concluding our investigation of engineering students, Chapter 6 supplements the grade-focused analyses of Chapters 4 and 5 by considering how the self-efficacy of engineering majors in their various foundational courses differs by gender. There is a significant underrepresentation of women in many Science, Technology, Engineering, and Mathematics (STEM) majors and careers. Prior research has shown that self-efficacy can be a critical factor in student learning, and that there is a tendency for women to have lower self-efficacy than men in STEM disciplines. This study investigates gender differences in the relationship between engineering students' self-efficacy and course grades in foundational courses. By focusing on engineering students, we examined these gender differences simultaneously in four STEM disciplines (mathematics, engineering, physics, and chemistry) among the same population. Using survey data collected longitudinally at three time points and course grade data from five cohorts of engineering students at a large US-based research university, effect sizes of gender differences are calculated using Cohen's d on two measures: responses to survey items on discipline-specific self-efficacy and course grades in all first-year foundational courses and second-year mathematics courses. In engineering, physics, and mathematics courses, we find sizeable discrepancies between self-efficacy and performance, with men appearing significantly more confident than women despite small or reverse direction differences in grades. In chemistry, women earn higher grades and have higher self-efficacy. The patterns are consistent across courses within each discipline. All self-efficacy gender differences close by the fourth year except physics self-efficacy. The disconnect between self-efficacy and course grades across subjects provides useful clues for targeted interventions to promote equitable learning environments. The most extreme disconnect occurs in physics and may help explain the severe underrepresentation of women in "physics-heavy" engineering disciplines, highlighting the importance of such interventions.

1.4 Enrollment Patterns in Physics in the Context of Other STEM Disciplines

Chapter 7 describes a study which situates achievement and enrollment patterns in physics within the larger context of these same patterns in other STEM disciplines. These investigations begin with an analysis of institutional data to understand the outcome of the many obstacles faced by students from historically disadvantaged backgrounds, which is an important avenue of research needed to work towards promoting equity and inclusion for all students. We use 10 years of institutional data at a large public research university to investigate the grades earned (both overall and in STEM courses only) by students categorized on four demographic characteristics: gender, race/ethnicity, low-income status, and first-generation college student status. We find that on average across all years of study and for all clusters of majors, underrepresented minority students experience a larger penalty to their mean overall and STEM GPA than even the most disadvantaged non-URM students. Moreover, the underrepresented minority students with additional disadvantages due to socioeconomic status or parental education level were even further penalized in their average GPA. Furthermore, we also find that while women in all demographic groups had a higher average overall GPA, these gender differences are almost completely non-existent in STEM GPA except among the most privileged students. These findings suggest that there is need to provide support to bridge the gaps that emanate from historical disadvantages to certain groups.

The preceding study in Chapter 7 cast a wide net to investigate how various demographic characteristics interacted with the academic achievement of students. Chapter 8 extends this by focusing on gender and considering how the enrollment patterns and grades of men and women compare in various STEM disciplines. Efforts to promote equity and inclusion using evidence-based approaches are vital to correct long-standing societal inequities that have disadvantaged women and discouraged them from pursuing studies, e.g., in many STEM disciplines. We use 10 years of institutional data at a large public university to investigate trends in the majors that men and women declare, drop after declaring, and earn degrees in as well as the GPA of the students who drop or earn a degree. We find that the majors with the lowest numbers of students also have the highest rates of attrition. Moreover, we find alarming GPA trends, e.g., women who drop majors on average earn higher grades than men who drop those majors, and in some STEM majors, women who drop the majors were earning comparable grades to men who persist in those majors. These quantitative findings call for a better understanding of the reasons students drop a major and for making learning environments equitable and inclusive.

Chapter 9 mirrors the study in Chapter 8 but instead focuses on the grades and persistence of students in different racial/ethnic groups. Underrepresented minority (URM) students are subjected to historically rooted inequities when pursuing an education, especially in STEM disciplines with little diversity. In order to make STEM education equitable and inclusive, evidence for how students from different racial/ethnic demographics are faring is necessary. We use 10 years of institutional data at a large public university to investigate trends in the majors that Asian, URM, and White students declare, drop after declaring, and earn degrees in as well as the GPAs of the students who drop or earn a degree. We find that higher percentages of the URM students drop most majors compared to other students and these trends are particularly pronounced in physics and economics. Moreover, we find alarming GPA trends in that the URM students consistently earn lower grades than their Asian and White peers. Furthermore, in some STEM disciplines, the URM students who earn a degree are earning the same grades as the Asian and White students who dropped the major. This troubling trend may signify lack of sufficient support, mentoring, and guidance to ensure excellence of the URM students who are already severely disadvantaged. These quantitative findings call for making learning environments equitable and inclusive so that many URM students who come to college with severe disadvantages are appropriately supported and can excel similar to other students.

Finally, in Chapter 10 we showcase a tool for self-paced online learning using "online learning modules" and the analysis thereof. We discuss earlier analysis efforts and their limitations, and subsequently show an example of an analysis using propensity score matching in order to obtain comparable subsamples of students across different years of study. The study examines which elements of student behavior may be useful for detecting different modes of engagement, and what effect those modes can have on subsequent analysis. In particular, we first consider the prevalence of students making brief initial attempts in order to gain access to instructional material. Second, we consider performance on a scaffolding "on-ramp" module to detect whether students are being reminded of prior knowledge or learning from the scaffolding problem. We find that after removing the roughly 50% of students who do not seriously engage with their first attempts from analysis, the inclusion of the on-ramp module primarily benefits those students who already possessed some mastery of the material. These results showcase the powerful avenues of analysis options afforded by clickstream data from online instructional materials, as well as the continued need to develop learning materials that serve to benefit all students. Such analyses can be an extremely powerful tool in order to peer into the student learning process and could be of great value in future studies of student learning and motivation in physics.

2.0 Physics Curriculum - Structural Equation Modeling

2.1 Introduction

Increasingly, universities are using evidence to improve student learning as well as to promote equity and inclusion so that students from diverse background and demographics can succeed [218, 66, 119, 118, 238, 78, 72, 17, 237]. Institutional data can play a critical role in understanding the successes and the areas where changes are needed in physics departments [288].

Holistic consideration of how physics departments are currently succeeding in supporting their undergraduate majors is crucial in order to make appropriate changes to the curricula and pedagogies for the majors based upon metrics informed by data and ensure that students are adequately supported and advised. These considerations include how prerequisite physics and mathematics courses predict performance in subsequent physics courses throughout the curriculum for physics majors, and such investigations are vital regardless of the theory of change [119, 72] a physics department adopts and implements based upon its institutional affordances and constraints. At the same time, with advances in digital technology in the past decade, data analytics can provide valuable information that can be useful in transforming learning for all students [6, 217].

This research is inspired by the fact that while many investigations in physics education have focused on evidence-based classroom practices to improve student learning at all levels in the physics undergraduate curriculum [184, 185, 248, 162, 290, 291, 241, 231, 105], there is significantly less focus on the connection between how student performance in different subsequent courses builds on prior courses in the physics curriculum overall. Information obtained from data analytics on large institutional data in these areas can be an important component of understanding, e.g., the role the earlier courses play in later physics course performance as well as contemplating strategies for strengthening these ties, and improving physics major advising, mentoring and support.

Moreover, it is important that physics departments take a careful look at the extent to which their programs for the majors are equitable and inclusive. It is imperative that we structure our programs to provide adequate support, advising and mentoring to all students, including women and students from diverse ethnic and racial backgrounds who have traditionally been left out [200, 199] in order to ensure that all students have sufficient opportunity to excel as a physics major. Further, studies of institutional data are being increasingly used to investigate issues of equity and inclusion, especially for women and underrepresented minorities in physics and Science, Technology, Engineering, and Mathematics (STEM) disciplines as a whole [168, 121, 181, 212, 166, 92, 172, 192, 167, 211, 283, 151, 171, 186, 83, 84, 143, 145, 144, 105, 1, 35, 122].

In this study we focus in particular on the issue of gender equity in physics education. There has been much important work done in recent years studying the many and various ways that women are not being adequately supported by their physics departments [17, 147, 2, 136, 137, 175, 174, 100, 265, 263, 177, 178, 264, 169, 142, 115, 117, 116, 232, 240, 143, 83, 145, 144, 84, 105, 1, 35, 122]. In particular, inequitable gender differences in introductory physics classes have been documented with a variety of measures including motivational characteristics [263, 177, 178, 142, 140, 207, 277, 141, 115, 143, 145, 144, 209, 139, 208, 116, 205, 173, 164], grades [178, 140, 207, 277, 141, 147, 2, 117, 231, 173], and performance on conceptual inventories [168, 17, 264, 169, 121, 175, 174]. The overarching theme of these studies' conclusions is in line with broader studies of gender inequity in physics and STEM [265, 151, 171, 181, 136, 137, 283, 1, 35, 122, 240] which highlight the significant obstacles faced by women in STEM which arise from societal stereotypes and biases. We seek to extend these critically important studies by looking at gender differences in the grade outcomes of physics majors not just in introductory physics but throughout a physics curriculum.

In order to gain an understanding of these issues central to improving physics education for the majors, this research harnesses data analytics in the context of a large state-related university to investigate how well the performance of physics majors in physics and mathematics courses throughout a physics curriculum predicts performance in subsequent physics courses. We note that the first-year physics and mathematics courses are very similar at most colleges in the US. Moreover, many of the advanced courses in these subjects also have well-defined curricula that are common across many colleges and universities. These courses for the majors have been offered over decades under the assumption that the later physics courses would build on the earlier ones coherently to help the majors build a robust knowledge structure of physics and develop their problem solving, reasoning and meta-cognitive skills.

Here we discuss an investigation that uses data analytics applied to 15 years of institutional data for physics majors to analyze not only these relationships between course performance in different years, but also whether there are any gender differences in these curricular relationships. Course grades are an important measure because in the short term they are the measure that students themselves will see and use to inform their attitudes and decisions about physics, while in the long-term they are the measure that will be reflected on their transcript and affect future career opportunities. Further, it is important to further our understanding of course grades since they are a consistently available measure in institutional data going back years or decades.

The investigation can be useful for other institutions who may perform similar analyses in order to contemplate strategies for improving education for physics majors in a holistic manner. In particular, institutions could compare their findings with the baseline data from a large state-related university presented here for the synergy observed between the required courses in the curriculum for the physics major.

2.2 Research Questions

Our research questions regarding the physics curriculum for the majors at a large state-related university are as follows.

RQ1. Where in the curriculum do gender differences in course performance occur for physics majors (i.e., in introductory and advanced physics and mathematics courses)?

RQ2. Does performance in introductory physics and mathematics courses predict performance in advanced physics and mathematics courses?

RQ3. Does the degree to which earlier course grades predict later course grades differ for men and women?

2.3 Methodology

2.3.1 Measures

Using the Carnegie classification system, the university at which this study was conducted is a public, high-research doctoral university, with balanced arts and sciences and professional schools, and a large, primarily residential undergraduate population that is full-time and reasonably selective with low transfer-in from other institutions [133]. De-identified data were provided by the university on all students who had enrolled in introductory physics from Fall 2005 through Spring 2019. The data include demographic information such as gender. We note that gender is not a binary construct. However, the university data includes "gender" as a binary categorical variable. Therefore, that is how the data regarding gender are represented in these analyses. From the full sample from 2005-2019, a sub-sample was obtained by applying several selection criteria to select out physics majors from those from other majors who took introductory physics. In particular, in order to be kept in the sample, students were required to 1) declare a physics major at any point or be a non-engineering student enrolled in the honors introductory sequence and 2) enroll in Modern Physics. Note that all of the courses we consider in this analysis in Table 1 are the required lecture courses in the curriculum for the physics major. We consider only required courses in order to maintain as consistent a population as possible. Further, we consider only lecture courses since the contemporary laboratory courses have very high and narrow grade distributions (with over 90% of students receiving an A) that are less suited for investigations of gender differences. After applying the selection criteria, the sample contains 451 students, which are 19.5% female and have the following race/ethnicities: 80.5% White, 10.9% Asian, 2.4% Latinx, 2.2% African American, and 3.8% Other or Unspecified.

The majority of the considered courses are taught in an active learning style, with the

Table 1: All required lecture courses in physics and mathematics taken by physics majors are listed. Full course names are given along with shortened names used elsewhere in this paper and the term(s) in which the courses are typically taken by physics majors.

Typical Term	Full Course Name	Shortened Course Name
1	Basic Physics for Science and Engineering 1	Physics 1
1	Basic Physics for Science and Engineering 1 - Honors	Honors Physics 1
2	Basic Physics for Science and Engineering 2	Physics 2
2	Basic Physics for Science and Engineering 2 - Honors	Honors Physics 2
3	Intro to Thermodynamics, Relativity and Quantum Theory	Modern Physics
5,6	Computational Methods in Physics	Comp. Methods
5,6	Electricity and Magnetism	EM
6,7	Mechanics	Mechanics
6,7	Thermodynamics and Statistical Mechanics	Thermo
7	Introduction to Quantum Mechanics 1	QM 1
1	Analytic Geometry and Calculus 1	Calc 1
1, 2	Analytic Geometry and Calculus 2	Calc 2
2, 3	Analytic Geometry and Calculus 3	Calc 3
4	Introduction to Matrices and Linear Algebra	Linear Algebra
5	Applied Differential Equations	Diff. Eq.

remainder taught in a traditional lecture style. We further note that multiple instructors have taught each course in the physics curriculum within the investigated time period, since the physics department has a policy that the same instructor cannot teach the same course for more than two years. However, in some courses such as the honors introductory sequence, those multiple instructors have all been men.

The data also include high school GPA on a weighted 0-5 scale that includes adjustments to the standard 0-4 scale for Advanced Placement and International Baccalaureate courses. This weighting was performed by the university prior to the data being supplied for research. Further, students' declared majors are recorded in the data separately for each term in which they are enrolled.

Finally, the data include the grade points and letter grades earned by students in each course taken at the university. Grade points are on a 0-4 scale with A = 4, B = 3, C = 2, D = 1, F = 0, where the suffixes '+' and '−' respectively add or subtract 0.25 grade points (e.g. B− = 2.75), with the exception of A+ which is reported as the maximum 4 grade points. In this study we consider course grades to be a proxy of student learning. The extent to which this assumption is true is not as important as the extent to which it aligns with how course grades are viewed by students and future employers or graduate schools. Students will use their course grades to inform decisions about their future, from their academic major to their career path.

2.3.2 Analysis

In order to evaluate the grades that the physics majors earn in physics and mathematics courses, we grouped students by the gender variable and computed standard descriptive statistics (mean, standard deviation, sample size) separately for each group. Gender differences in course grades were initially evaluated using Cohen's d to measure the effect size [201, 195], as is common in education research [206].

The extent to which performance (i.e., grades earned) in earlier physics and mathematics courses predicts performance in later physics and mathematics courses was evaluated using Structural Equation Modeling (SEM)[152]. SEM is the union of two statistical modeling techniques, namely Confirmatory Factor Analysis (CFA) and Path Analysis. The CFA portion tests a model in which observed variables (or "indicators") are grouped into latent variables (or "factors"), constructed variables that represent the variance shared among all indicators that load on that particular factors. The degree to which each indicator is explained by the factor is measured by the standardized factor loadings, λ (with $0 \leq \lambda \leq 1$), where λ^2 gives the percentage of variance in the indicator explained by the factor. The Path Analysis portion then tests for the statistical significance and strength of regression paths between these factors, simultaneously estimating all regression coefficients, β, throughout the model. This is an improvement over a multiple linear regression model in which only a single response (target or outcome) variable can be predicted at a time, which problematically disallows hierarchical structures [267]. By estimating all regression paths simultaneously, all estimates are able to be standardized simultaneously, allowing for direct comparison between standardized β coefficients throughout the model.

In this paper, we report the model fit for SEM using the Comparitive Fit Index (CFI), Tucker-Lewis Index (TLI), and Root Mean Square Error of Approximation (RMSEA). Commonly cited standards for goodness of fit using these indices are as follows: For CFI and TLI, Hu and Bentler [128] found that many authors [128, 51, 229] suggest values above 0.90 and 0.95 indicate a good fit and a great fit, respectively. For RMSEA, several authors [128, 48] suggest that values below 0.10, 0.08, and 0.05 indicate a mediocre, good, and great fit, respectively.

Finally, these model estimations can be performed separately for different groups of students (e.g., men and women) using multi-group SEM. These differences are measured in a series of tests corresponding to different levels of "measurement invariance" in the model [152], with each step fixing different elements of the model to equality across the groups and comparing to the previous step via a Likelihood Ratio Test (LRT). A non-significant p-value at each step indicates that the estimates are not statistically significantly different across groups. "Weak" measurement invariance is demonstrated by fixing to equality the factor loadings, "strong" invariance is demonstrated by further fixing to equality the indicator intercepts, and finally "strict" invariance is demonstrated by further fixing to equality the residual error variance of the indicators. If measurement invariance holds, then all remaining differences between the groups occur at the factor level, either as differences in factor intercepts or β coefficients. Further, if no differences are found in β coefficients, then any remaining group differences in factor intercepts may be modeled by including a categorical grouping variable which directly predicts the factors.

Using SEM, we model student progression through the physics curriculum by grouping courses together into factors by their subject (physics or mathematics) and the order in which the courses are typically taken by physics majors. In particular, we group introductory courses taken within the first year together (Physics 1 and 2 as "Intro Physics" and Calculus 1 and 2 as "Intro Math"), and consider the remaining courses beyond the first year separately. This separation between first-year and other courses is designed to test our hypothesis that performance in introductory physics and advanced physics courses are different constructs, and the relationship between these two factors in an SEM model will determine how closely related they are, controlling for performance in mathematics courses. We use multi-group SEM to test for gender moderation, i.e., to test for gender differences in the model, including mean differences of courses (indicators) and course factors. Since we found no gender differences anywhere except in factor-level intercepts, we ultimately model the gender differences not with multi-group SEM, but with a categorical "Gender" variable directly predicting items with different intercepts.

Due to the nature of institutional grade data, modeling students' progress through an entire curriculum involves a large amount of missing data due to various factors. These can include students receiving credit for courses taken elsewhere (e.g., over the summer at a different college), not completing the curriculum, skipping courses that are normally required with special permission, and the inevitable errors that occur in large datasets. We note that due to strict requirements by the department, very few students receive credit for Physics 1 from Advanced Placement or International Baccalaureate, and no students can receive such credit for Physics 2. The default approach to missing data in many modeling programs, listwise deletion, is then not desirable since it leaves very few students in the sample and can bias the results [204]. Considering this, we employed Full Information Maximum Likelihood (FIML) using the R package `lavaan` [234] in order to impute missing data within the SEM model [152].

In addition to the aforementioned benefits of using SEM such as simultaneous estimation of all model elements and the ability to use FIML for missing data estimation, the basic structure of SEM also provides benefits to the modeling process. In particular, by first using CFA to group indicators into factors and then performing path analysis on those factors, the effect of measurement error is minimized since the error variance will be left at the indicator level and does not contribute to the estimation of regression coefficients at the factor level [152].

All analyses were conducted using R [226], making use of the package `lavaan` [234] for the SEM analysis and the package `tidyverse` [279] for data manipulation and descriptive statistics.

2.4 Results

In order to investigate for gender differences in course grades and answer **RQ1**, we grouped students by the gender variable and first calculated the standardized mean difference, Cohen's d, to measure the effect size of the gender differences [201, 195]. Table 2 shows these results for the required physics and mathematics courses for prospective physics majors in their first year courses, regardless of whether they continued on in the curriculum, while Table 3 shows these results for only those who at least continued through Modern Physics. Though all later analyses are performed on the student population shown in Table 3, namely those physics majors who persist at least through the second year, the contrast between those students and the first-year prospective physics in Table 2 shows that on average higher than average performing women in the Honors Physics courses are switching out of a physics major. Note that in Table 3 some courses have lower N than Modern Physics for a variety of reasons such as skipping the course with Advanced Placement credit (Physics 1, Calculus 1, Calculus 2), the course not always being required for the major (Comp. Methods, Thermo, QM 1), or students either having requirements waived or obtaining credit at other universities (potentially all courses). Moreover, there is significant attrition in the number of students between Table 2 and Table 3. This attrition is among students who declared the major, which is certainly an underestimation since not every prospective physics major will declare the major before dropping. This is a trend which could be investigated in a future study in order to contextualize the high attrition rate in physics by comparing attrition rates among different STEM majors.

We find that, on average, men performed slightly better than women in all introductory physics courses, with Cohen's d ranging from -0.12 to -0.24 among all prospective physics majors (Table 2), indicating a small effect size, and ranging from -0.16 to -0.49 for those who continue to Modern Physics (Table 3), indicating a small to medium effect size. Gender differences in mathematics and advanced physics courses (Table 3) are mixed, with no clear pattern of performance differences.

The statistical significance of these gender differences is first tested using a multivariate analysis of variance (MANOVA) on three clusters of courses in Table 3, namely introductory physics, advanced physics, and mathematics. Courses were clustered in order to keep the number of students in the MANOVA from dropping too low, since MANOVA employs listwise deletion. These results support the patterns noted before: that introductory physics ($F(2, 371) = 3.13$, $p = 0.045$) displays a consistent pattern of men earning higher grades than women, albeit only marginally significant at the $p < 0.05$ level with the listwise deletion employed by MANOVA. Further, there is no consistent pattern in either advanced physics ($F(6, 119) = 1.52$, $p = 0.179$) or advanced mathematics ($F(5, 108) = 1.07$, $p = 0.379$), evidenced by $p > 0.05$ for each of these tests. A more sophisticated test of these gender differences will occur in the investigation of **RQ3**, where we can use multi-group SEM to test for gender differences among all elements of the model, including differences in the means earned by men and women in each course. In addition, multi-group SEM allows us to perform these tests while using FIML to estimate missing data, a significant improvement over listwise deletion.

The full grade distributions can be found in Fig. A.1 and Fig. A.1 in Appendix A.

Turning then to **RQ2**, we use SEM to test for the degree to which performance in earlier courses predicts that of later courses in the curriculum. The full 451 student sample was used in all SEM models, with FIML employed to impute missing data. We grouped courses into four broad categories: introductory physics (with the regular and honors sequences combined), advanced physics (all physics beyond the introductory sequence), and introductory mathematics (Calculus 1 and Calculus 2), and allowed regression paths forward in time from introductory to advanced courses.

The final model is shown in Fig. 1 (CFI = 0.947, TLI = 0.933, RMSEA = 0.053), in which non-significant regression paths have been trimmed from the model. One notable feature of Fig. 1 is that introductory mathematics strongly predicts advanced mathematics, as expected, which covaries strongly with advanced physics. However, introductory physics does not predict advanced physics at all while introductory mathematics does, indicating that the primary predictor of success in advanced physics courses is success in mathematics courses.

Figure 1 is not the only possible model for the relationships among courses. In particular, the majority of students take all of the advanced mathematics courses either before or concurrently with all advanced physics courses beyond Modern Physics. A model in which advanced mathematics predicts rather than covaries with advanced physics is shown in Fig. 2 (CFI = 0.946, TLI = 0.934, RMSEA = 0.053). Yet another model is shown in Fig. 3 (CFI = 0.950, TLI = 0.936, RMSEA = 0.052), in which the advanced physics factor has been split according to the typical time-order in which students take the courses. No models tested show introductory physics predicting advanced physics when controlling for introductory and/or advanced mathematics, including those not shown here such as a model in which introductory mathematics is allowed to predict introductory physics, rather than covary with it.

To test for gender differences and answer **RQ3**, we first used multi-group SEM to estimate the model separately for men and women, and then used a series of likelihood ratio tests to test for differences in the model [152], first testing factor loadings, then indicator intercepts, then residual variances, then finally regression paths. In each step, the model fit was moderate to good, with CFI > 0.90, TLI > 0.90, and RMSEA < 0.08. Each step produced statistically non-significant changes from the previous according to LRTs, indicating that the estimates could be fixed to equality across the two groups ($p > 0.10$ for each step). The only statistically significant gender differences occurred in the intercepts of high school GPA, which is not an indicator for any factor, and the introductory physics factor.

Since there were no statistically significant gender differences in regression coefficients, we converted the model from a multi-group SEM to a model that includes gender as a binary categorical variable (1 for "F" and 0 for "M") predicting high school GPA and introductory physics.

In all three models (Figs. 1, 2, and 3) the gender differences take on the same form: on average, women have slightly higher high school GPA ($\beta = 0.14$, $p = 0.002$), while men are predicted to have slightly higher grades in introductory physics ($\beta = -0.19$, $p < 0.001$) when controlling for the high school GPA difference, and no other gender differences are predicted anywhere else in the model. To expand further, the statistically significant path from gender to introductory physics means that men are predicted to have higher grades in introductory physics than women with the same high school GPA. For the courses other than introductory physics, this means that the inconsistent gender differences observed in mathematics and advanced physics courses in Table 3 are statistically non-significant when controlling for high school GPA, which either directly or indirectly predicts every other course in the model.

2.5 Discussion

In answering each of the research questions, the introductory physics sequence stood out as behaving differently from the other courses, and the overall picture paints the introductory sequence as the only gender-imbalanced part of the entire physics curriculum (pertaining to differential performance of men and women). In particular, answering **RQ1**, Tables 2 and 3 together with the gender differences observed in the SEM models in Figs. 1, 2, and 3 show that introductory physics courses are the only ones in the curriculum with statistically significant gender differences, with men earning higher grades on average than women. The SEM models provide further context, showing that all other gender differences are non-significant when controlling for high school GPA, which is higher on average for female physics majors than their male counterparts. Thus, even though men only earn higher grades in introductory physics with a small effect size, that small effect size is slightly larger in magnitude and opposite in sign to the effect size of women's higher average high school GPA. One hypothesis for why there is gender difference in performance in introductory courses is that those courses are taken in the first year in large-enrollment classes. Due to societal stereotypes and biases associated with physics, women may have a lower sense of belonging and self-efficacy [8, 9, 10] in those types of impersonal, non-equitable, and non-inclusive learning environments which can impact learning [254, 253, 18, 263, 177, 178, 142, 115, 117, 116, 232, 240, 143, 83, 145, 144, 84].

Even though the observed gender differences only occur in the introductory courses, this situation is pernicious and deeply troubling. The first experience of these women in first-year physics is in courses where they are faced with enormous pressure from societal stereotypes [232, 265, 136, 137, 263, 177, 178, 142, 115, 143, 83, 145, 144, 84, 105, 1, 35, 122] that results in course performances inconsistent with their experiences in high school as well as concurrent experiences in their mathematics courses. To be precise, the women in these introductory physics courses will be given inaccurate data about their ability to succeed in physics during the time in which they are making crucial decisions about their future. This situation is extremely problematic, only serves to perpetuate the societal stereotypes and biases that puts an undue burden on these women to begin with [232, 265, 136, 137, 263, 177, 178, 142, 115, 143, 83, 145, 144, 84, 105, 1, 35, 122].

In answering **RQ2**, the SEM model in Fig. 1 shows that performance in introductory physics does not predict future grades earned in advanced physics courses when controlling for performance in introductory mathematics, and this is true for both men and women. We also note that whether we allow advanced mathematics to covary with advanced physics (Fig. 1) or predict advanced physics directly (Fig. 2), we find no statistically significant regression path from introductory to advanced physics. However, allowing advanced mathematics to predict (via a regression path) rather than covary with advanced physics leads to advanced physics being predicted solely by advanced mathematics (and not by introductory mathematics). That is, in Fig. 2, introductory mathematics strongly predicts advanced mathematics, which in turn strongly predicts advanced physics. One reason for why introductory mathematics only predicts advanced physics via advanced mathematics (Fig. 2) is that the content of Calculus 1 and 2 courses (e.g., evaluating limits and simple differentiation and integration) is less directly relevant to success in advanced physics courses. While one is expected to know simple differentiation and integration in advanced physics courses, most of the variance in advanced physics performance is due to student proficiency in vector calculus, linear algebra and differential equations (in fact, in these physics courses, students generally get full credit for leaving the final answer as an integral if the limits and integrand are correct).

Further, Fig. 3 explores the relationship among future physics courses and finds statistically significant regression paths from Modern Physics (the sole required 2nd year physics course) to 3rd year physics to 4th year physics, even when controlling for advanced mathematics. Yet still, Fig. 3 shows no connection from introductory physics to any other courses. This makes the lack of a connection from introductory physics to future physics courses unique in the physics sequence.

Considering our findings from **RQ1**, this lack of a connection between introductory and advanced physics makes the situation even more problematic. The inaccurate feedback that women are receiving in introductory physics, which is pressuring them away from pursuing physics majors, does not have any predictive power for how well these women could succeed in advanced physics courses.

One hypothesis for why only advanced mathematics courses predict performance in advanced physics courses while introductory physics courses do not is that advanced physics courses essentially test student facility with mathematical procedures as opposed to their conceptual understanding which is typically the focus in introductory physics courses. In particular, students can typically do very well in advanced physics courses if they have just enough knowledge of advanced physics in order to recognize which mathematical procedure to use (e.g., solving a boundary value problem) even if their conceptual foundation in physics is weak (which is the focus of introductory physics). The lack of gender differences in all mathematics courses as well as the advanced physics courses supports this hypothesis, indicating that there is a fundamental difference in how introductory physics courses are taught. In fact, our earlier investigation pertaining to conductors and insulators suggests that advanced physics students on average do not perform better on conceptual questions at the level of introductory physics than introductory physics students [32]. Moreover, in another investigation, many students in advanced graduate courses did not perform significantly better than introductory students and admitted that they had no time to think about concepts and were essentially solving mathematics problems without learning physics from their advanced courses [250].

Finally turning towards **RQ3**, we find that in all three SEM models tested (Figs. 1, 2, and 3), there were no significant gender differences in any predictive paths in the model. The gender differences only occurred in two places: the intercepts of high school GPA (higher on average for women) and the introductory physics factor (higher on average for men). Introductory physics did not predict forward at all in the SEM model, but high school GPA predicts every course factor in the model either directly or indirectly. This means that any gender differences elsewhere in the model (i.e., not in introductory physics) are consistent with those observed in high school GPA.

We note that this model focuses on the relationships between the grades earned and does not account for other ways in which gender disparities in introductory physics can affect students (e.g., through self-efficacy, sense of belonging, etc.). However, grades earned play a key role in students' crucial decisions about whether to remain in college and which major to pursue [24, 274, 161, 292]. In particular, one mechanism by which this occurs is the feedback loop between course grades and self-efficacy [292, 8, 9, 10, 224, 214, 46]. Other studies at this same university have found significant gender differences favoring men in the physics self-efficacy of students in large introductory physics courses [177, 178, 144, 145, 140, 141, 277], consistent with studies at other universities [53, 239]. Although our analysis only includes students who had not only declared a physics major but had also completed the modern physics course in the sophomore year, the gender differences in grades earned in introductory physics courses could have a large, gender differential impact on students' choice to pursue physics and related majors, despite the fact that performance in introductory physics does not predict performance in advanced physics courses. These findings further suggest the need for efforts towards improving equity and inclusion in introductory physics courses, including interventions designed to boost students' self-efficacy, growth mindset and sense of belonging in physics [280, 270, 271, 33].

In conclusion, a completely cohesive curriculum for physics majors should not only be consistent in academic content from year to year, but also in its positive and inclusive environment so that students from all demographics can excel including those groups which have traditionally been underrepresented in physics. We urge researchers at other institutions to perform similar analyses in order to evaluate the efficacy of the assumptions underlying the curriculum for physics majors, and how well the various courses required for physics majors cohere. Furthermore, it is critical that other physics departments investigate gender differences within their curricula in order to counteract situations such as the one presented here, where women in introductory courses may be using inaccurate data about their performance in physics to inform decisions about their future. The situation observed in this study indicates that actions to remedy these pernicious gender differences are imperative, and it is crucial that all physics departments make every effort possible to seek and stamp out similar trends at their own institutions. Ignoring such trends only serves to perpetuate the biases and stereotypes that disproportionately affect women in physics.

2.6 Acknowledgments

This research is supported by the National Science Foundation Grant DUE-1524575 and the Sloan Foundation Grant G-2018-11183. We thank Robert Devaty for his invaluable feedback on this manuscript.

Table 2: Descriptive statistics are reported for prospective physics majors in introductory physics courses. To be included in this table, students need only have declared a physics major, not necessarily enrolled in advanced courses, in order to briefly examine all students who declared a physics major during their first year. The reported statistics include the sample size (N), mean grade points earned (μ), and standard deviation of grade points (σ) in each course for men and women separately, along with Cohen's d measuring the effect size of the gender difference. $d < 0$ indicates the mean for men is higher, $d > 0$ indicates the mean for women is higher.

Course	Gender	N	μ	σ	d
Physics 1	F	64	2.87	0.92	-0.24
	M	238	3.08	0.83	
Honors Physics 1	F	71	3.20	0.63	-0.12
	M	221	3.29	0.83	
Physics 2	F	80	2.76	1.02	-0.15
	M	292	2.91	0.95	
Honors Physics 2	F	59	3.27	0.61	-0.18
	M	195	3.39	0.70	

Table 3: Descriptive statistics are reported for physics and mathematics courses taken by physics majors who have at least taken physics courses up to and including Modern Physics. Reported are the sample size (N), mean grade points earned (μ), and standard deviation of grade points (σ) in each course for men and women separately, along with Cohen's d measuring the effect size of the gender difference. $d < 0$ indicates the mean for men is higher, $d > 0$ indicates the mean for women is higher. Three multivariate analyses of variance (MANOVA) are reported, with courses grouped to reduce listwise deletion into introductory physics, advanced physics, and mathematics.

Course	Gender	N	μ	σ	d	MANOVA
Physics 1	F	44	2.99	0.74	-0.25	$F(2,371) = 3.13$, $p = 0.045$
	M	178	3.17	0.73		
Honors Physics 1	F	35	3.13	0.67	-0.49	
	M	124	3.45	0.66		
Physics 2	F	55	2.93	0.91	-0.16	
	M	221	3.07	0.84		
Honors Physics 2	F	32	3.24	0.71	-0.26	
	M	124	3.42	0.67		
Modern Physics	F	88	3.03	0.75	0.02	$F(6,119) = 1.52$, $p = 0.179$
	M	363	3.01	0.98		
Comp. Methods	F	43	3.15	1.05	0.12	
	M	170	3.01	1.19		
EM	F	56	2.74	0.84	0.07	
	M	223	2.67	1.08		
Mechanics	F	57	2.65	0.84	-0.33	
	M	215	2.94	0.87		
Thermo	F	42	2.92	0.96	0.04	
	M	160	2.88	1.02		
QM 1	F	33	3.27	0.92	-0.05	
	M	139	3.31	0.89		
Calculus 1	F	34	3.15	0.85	0.11	$F(5,108) = 1.07$, $p = 0.379$
	M	149	3.05	0.99		
Calculus 2	F	59	2.77	0.97	-0.05	
	M	232	2.82	1.08		
Calculus 3	F	84	2.96	1.02	0.07	
	M	335	2.88	1.15		
Linear Algebra	F	56	3.15	0.87	0.21	
	M	230	2.93	1.18		
Diff. Eq.	F	51	2.58	1.06	-0.21	
	M	236	2.82	1.18		

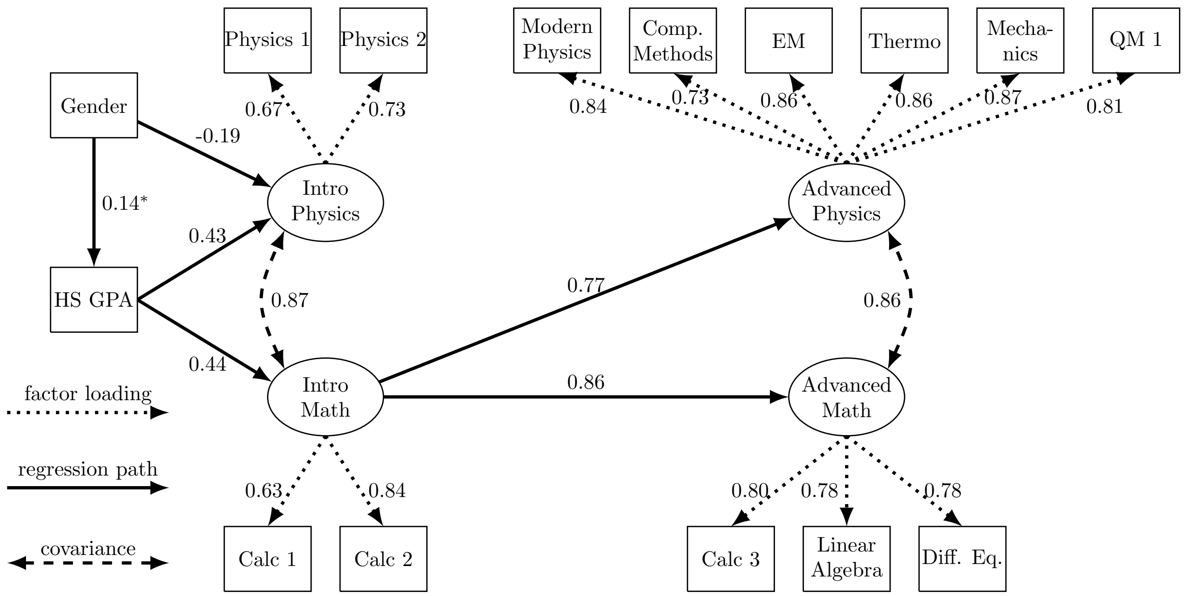


Figure 1: A diagram of the SEM model designed to test for the relationship between physics and mathematics courses in the physics curriculum, as well as gender differences therein. Reported next to each line are the standardized values for factor loadings, regression coefficients, and covariances. The gender variable was coded as 1 for "F" and 0 for "M", so paths from gender with $\beta > 0$ and $\beta < 0$ indicate a higher mean for women and men, respectively, in the predicted variable. All drawn paths are significant to the $p < 0.001$ level except those denoted with a superscript $*$, which are significant to the $p < 0.01$ level. All missing paths are not statistically significant, with $p > 0.05$.

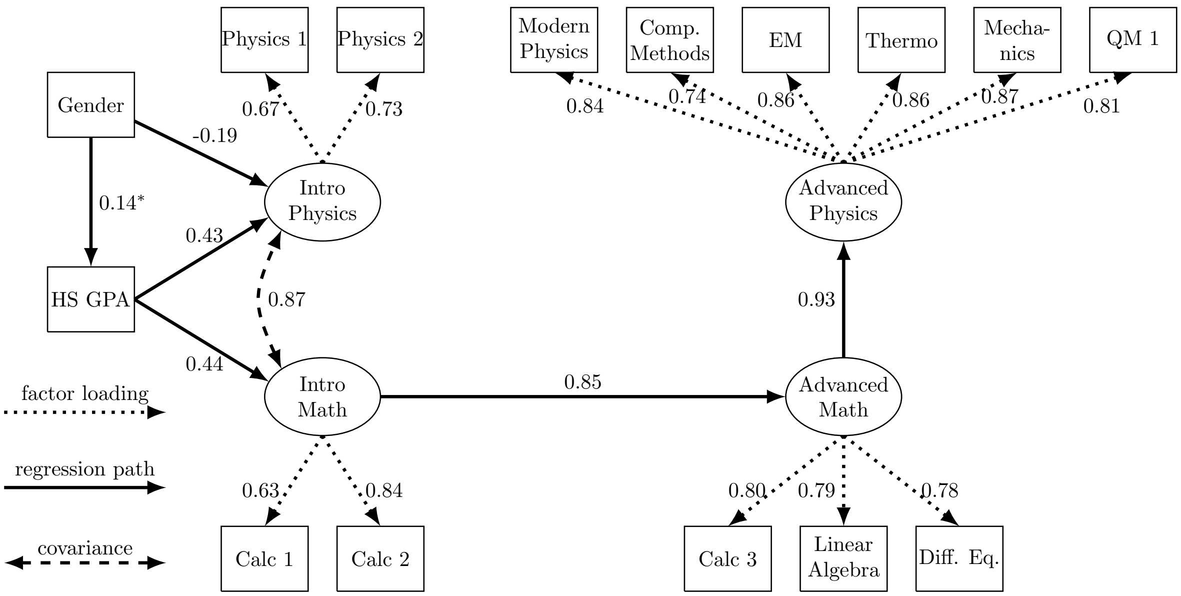


Figure 2: An alternate model to the one shown in Fig. 1, with the Advanced Math factor allowed to predict Advanced Physics. Reported next to each line are the standardized values for factor loadings, regression coefficients, and covariances. The gender variable was coded as 1 for "F" and 0 for "M", so paths from gender with $\beta > 0$ and $\beta < 0$ indicate a higher mean for women and men, respectively, in the predicted variable. All drawn paths are significant to the $p < 0.001$ level except the one denoted with a superscript $*$, which is significant to the $p < 0.01$ level. All missing paths are not statistically significant, with $p > 0.05$.

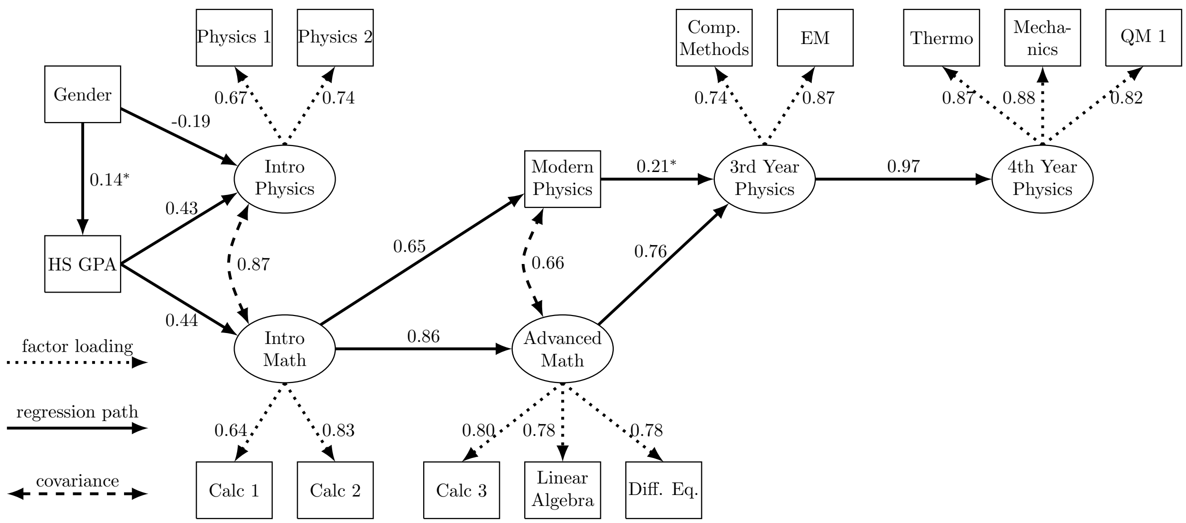


Figure 3: An alternate model to the one shown in Fig. 1, with the Advanced Physics factor split by the year in which the courses are typically taken. Reported next to each line are the standardized values for factor loadings, regression coefficients, and covariances. The gender variable was coded as 1 for “F” and 0 for “M”, so paths from gender with $\beta > 0$ and $\beta < 0$ indicate a higher mean for women and men, respectively, in the predicted variable. All drawn paths are significant to the $p < 0.001$ level except the one denoted with a superscript $*$, which is significant to the $p < 0.01$ level. All missing paths are not statistically significant, with $p > 0.05$.

3.0 Recognition Always Matters: A Cross-Sectional Study of the Physics Identity of Physics Majors

3.1 Introduction

In the past few decades, physics education researchers have conducted diverse investigations focused on improving student understanding [74, 76, 83, 90, 97, 148, 159, 179, 269, 275]. Moreover, in recent years there have been several studies building a framework of physics identity [116, 114, 142, 143]. It is important to investigate physics identity since many studies have found that physics identity can play a major role in the academic trajectory and future career choices of undergraduates [116, 114, 142, 143, 183, 164, 73, 102]. First, Hazari and colleagues [116, 115] developed a framework for physics identity based on the earlier work of Carlone and Johnson [50] that stated that a person's science identity depends upon performance feedback (e.g., grades), perceived recognition, and competency belief. The Hazari *et al.* framework [116, 115, 142, 143, 145] expands the work of Carlone and Johnson [50] by exploring physics identities in particular, and proposes that students' identity as a "physics person" is informed by these three factors (performance, perceived recognition, and competency belief) as well as by students' interest in physics. Finally, this framework was refined by Kalender *et al.* [142, 143, 145], who propose that, for physics identity, performance and self-efficacy (a concept similar to competency belief) are redundant predictors of physics identity. Thus, the Kalender *et al.* [142, 143, 145] framework for physics identity that we will test in this paper states that physics identity depends on three components:

- perceived recognition by instructors, TAs, peers, and family (or "external identity"),
- physics self-efficacy, and
- interest in physics.

Prior studies which focus on introductory physics students have found that perceived recognition is the strongest predictor of physics identity, which implies that physics instructors can boost their students' identification as a physics person by making an effort to recognize students as people who can excel in physics [116, 115, 142, 143, 145]. Self-efficacy, defined by Bandura as one's belief in one's ability to succeed at a particular task or subject [8, 9, 10, 11, 12, 13], is another extremely important motivational construct in this framework. Many prior studies of self-efficacy have shown that it is linked to many aspects of education, both in general and in physics specifically, including academic achievement, persistence in the major, and career choices after graduation [8, 9, 10, 11, 12, 13, 254, 253, 18, 145, 144, 177, 178, 214, 292, 24, 46, 157, 176, 272].

However, the population in introductory physics courses is primarily non-physics majors who are taking the course as a requirement for another degree. In order to understand how the framework of physics identity [116, 115, 142, 143, 145] applies to physics majors, we extend the previous studies to physics majors ranging from first-year undergraduate students to first-year graduate students by investigating responses on a validated survey over a five-year period. We further investigated how two additional aspects of students' attitudes about physics, namely, sense of belonging in the physics learning environment and student perception of their interactions with their peers, impacted their self-efficacy in physics. Student attitudes relating to the learning environment have been previously studied and shown to impact how students engage with physics [76, 83, 179, 275]. Further, important prior studies of the effect of the learning environment on students have thus far focused on introductory physics courses and revealed, e.g., many problematic gender trends [147, 74, 83].

Our research questions to guide this cross-sectional investigation are as follows.

RQ1. What are the trends over time for physics majors' self-efficacy, perceived recognition, interest, identity as a physics person, positive peer interaction, and sense of belonging?

RQ2. How do the responses differ for physics majors and non-majors in the same first-year courses?

RQ3. How consistent are the relationships between internal identity and the other motivational constructs for physics majors throughout their education?

RQ4. What are the relationships between self-efficacy, perceived recognition, and other constructs related to the classroom environment (perception of peer interaction and sense of belonging), and how do they evolve over time?

3.2 Methodology

Using the Carnegie classification system, the US-based university at which this study was conducted is a public, high-research doctoral university, with balanced arts and sciences and professional schools, and a large, primarily residential undergraduate population that is full-time and reasonably selective with low transfer-in from other schools [133].

3.2.1 Motivational Survey

A motivational survey was administered during the first and last two weeks of the semester in lecture courses throughout the undergraduate physics curriculum as well as first-year graduate courses. We will refer to the survey administered in the first two weeks of the semester as the "pre survey" (i.e., pre-instruction) and the survey administered in the final two weeks as the "post survey" (i.e., post-instruction). The survey was first administered in Fall 2015, with additions to the survey in Fall 2017 including items intended to measure students' physics identity, perceived recognition, perceived benefit of peer interaction, and sense of belonging in the physics learning environment. The survey took approximately 10 minutes for students to complete. Each course in which the survey was administered was taught by at least two different instructors during the collection period. Since we are interested in the motivational characteristics of physics majors, we used de-identified data provided by the university in order to determine which students were physics majors by restricting to those students who had, at any point in their education, declared a major in physics.

The survey items pertaining to this study, which were adapted from other validated surveys, are shown in Table 4 along with validation measures specifically for the physics majors (which will be discussed in the following section). Students responded to each item with the degree to which they agreed or disagreed with the statements on a Likert scale from 1–4 (1 being "strongly disagree" and 4 being "strongly agree"), with the exception of the first two items for "Interest" which had unique response options listed in the caption of Table 4 and the set of items relating to "Sense of Belonging" for which responses were on a 1–5 Likert scale. Prior to analysis, the responses to certain questions (indicated with a † in

Table 4) were reverse coded in order to align the Likert scale values.

3.2.2 Survey Validation

In previous studies, both the original and revised version of this survey were validated for calculus-based introductory physics courses, in which the majority of students are non-physics majors [177, 142, 143, 145, 144]. Since our study is primarily focused on physics majors, we validated the survey using only the responses of physics majors and found very similar results to those from the previous studies. First, interviews were conducted with 20 physics majors across all years of study which confirmed that the students were correctly interpreting the survey items. Next, we conducted an exploratory factor analysis (EFA) and found that the items group into the constructs labeled in Table 4 in a manner very similar to the previous studies [177, 142, 143, 145, 144]. A follow-up confirmatory factor analysis (CFA) found that the specified factors based upon the EFA have "good" model fit (SRMR < 0.08, CFI > 0.90, TLI > 0.90) [152, 48, 229, 128]. The standardized factor loadings (λ) from this CFA are reported for each item in Table 4, where λ^2 indicates the amount of variance explained in each item by the factors [152]. We also report Cronbach's α in Table 4 for each factor with multiple items. Finally, Table 5 contains the matrix of Pearson correlation coefficients between the identified factors [98, 152]. With the correlations all falling between 0.30 and 0.77, we find a good balance between these factors relating to one another without being so correlated that they are measuring the same construct.

3.2.3 Survey Respondents

Survey responses were grouped into "years" corresponding to the typical time line of physics courses for students at the studied university. For example, responses from students in introductory physics 1 (taken in the first year of study) are categorized as "year 1" while responses from students in modern physics (taken in the second year of study alongside several required mathematics courses) are categorized as "year 2." A majority of required physics courses for the majors – classical mechanics (CM), electricity and magnetism (EM), thermal physics beyond the introductory level – are taken in the third and fourth years and

Table 4: The survey items used in this study along with the response options given to students grouped by the associated motivational construct. For each construct with multiple items, Cronbach's α is provided as a measure of the internal consistency of these items. For each item, the standardized factor loading λ from a CFA model is listed, with three λ values specified for physics identity loading on each of the interest (λ_{Int}), perceived recognition (λ_{Rec}), and self-efficacy (λ_{SE}) factors. All items are on a Likert scale from 1 to 4 (1 = strong disagreement, 4 = strong agreement), except where otherwise indicated. Option set A = {Never, Once a week, Once a month, Every day}. Option set B = {Very boring, Boring, Interesting, Very interesting}. †Responses are reverse-coded prior to analysis in order to align the scales of all items. *Responses are on a 1–5 instead of 1–4 scale Likert scale.

Motivational Construct	α	λ	Question Text
Interest	0.82	0.52	• I wonder about how physics works [option set A]
		0.69	• In general, I find physics [option set B]
		0.73	• I want to know everything I can about physics
		0.76	• I am curious about recent discoveries in physics
		0.82	• I want to know about the current research that physicists are doing
Perceived Recognition	0.79	0.90	• My family sees me as a physics person
		0.90	• My friends see me as a physics person
		0.58	• My physics instructor/TA sees me as a physics person
Self-Efficacy	0.82	0.56	• †Other people understand more than I do about what is going on in this physics course
		0.65	• I am able to help my classmates with physics in the laboratory or in recitation
		0.68	• †I get a sinking feeling when I think of trying to tackle tough physics problems
		0.70	• I understand concepts I have studied in physics
		0.68	• If I wanted to, I could be good at physics research
		0.78	• If I study, I will do well on a physics test
		0.73	• If I encounter a setback in a physics exam, I can overcome it
Physics Identity	N/A	$\lambda_{\text{Int}} = 0.24$ $\lambda_{\text{Rec}} = 0.27$ $\lambda_{\text{SE}} = 0.46$	• I see myself as a physics person
Peer Interaction	0.94		My experiences and interactions with other students in this class...
		0.84	• Made me feel more relaxed about learning physics
		0.94	• Increased my confidence in my ability to do physics
		0.95	• Increased my confidence that I can succeed in physics
		0.91	• Increased my confidence in my ability to handle difficult physics problems
Sense of Belonging	0.87	0.83	• *I feel like I belong in this physics class
		0.78	• *†I feel like an outsider in this physics class
		0.84	• *I feel comfortable in this physics class
		0.68	• *I feel like I can be myself in this physics class
		0.74	• *†Sometimes I worry that I do not belong in this physics class

Table 5: The Pearson correlation coefficients (r) between the factors for each motivational construct in the CFA. Entries above the diagonal are omitted since the matrix is symmetric.

	Interest	Perceived Recognition	Self-Efficacy	Peer Interaction	Sense of Belonging
Interest	1.00				
Perceived Recognition	0.55	1.00			
Self-Efficacy	0.55	0.56	1.00		
Peer Interaction	0.33	0.30	0.54	1.00	
Sense of Belonging	0.42	0.45	0.77	0.65	1.00

although it is recommended that students take CM and EM courses before other upper-level courses, many students do not follow this recommendation. However, for the purposes of this study we group courses by the recommended curriculum and so responses from students in CM and EM courses are categorized as "year 3" and responses from students in quantum mechanics and thermal physics are categorized as "year 4". The sample size varies from year to year and in some cases differs between motivational constructs as items relating to physics identity and perceived recognition were added in Fall 2017. For matched pre and post responses, the sample size for physics majors ranges from 29 to 74, while unmatched post sample sizes range from 58 to 108 (in the third and fourth years, students took the survey in the fall and spring semesters so the number of responses is roughly twice those in other years). For non-physics majors the matched sample size ranges from 1314 to 1910 and the unmatched post sample size is 2109. We note that in our cross-sectional analysis, some undergraduate students may be included in multiple courses in the data over different years. However, the undergraduate and graduate student cohorts, though at the same university, are almost entirely separate. The exact sample size for each analysis will be reported along with the results.

3.2.4 Analysis

In order to investigate trends over time and answer **RQ1** and **RQ2**, we restricted the sample to matched pre and post responses (i.e., only those students who responded to both the pre and post survey items). Descriptive statistics – sample size, mean, and standard error – were calculated on these matched responses [98]. Then, in order to test the framework for physics identity and answer **RQ3**, we conducted multiple linear regression models [201, 195] on only the post survey responses from students in each year, as has been done in previous studies of this framework [116, 115, 142, 143, 145]. Since pre survey responses do not appear in the models, we did not restrict to the matched responses in these multiple linear regression models, instead using all available post survey responses (although the results are qualitatively similar using post survey responses for matched data).

All analyses were conducted using R [226], making extensive use of the package `lavaan` [234] for factor analyses and regression models and the package `tidyverse` [279] throughout all analyses and for data visualization.

3.3 Results

In order to answer **RQ1** and investigate the trends of physics majors' identity, perceived recognition, self-efficacy, and interest over time, we calculated the mean and standard error of each construct for physics majors in each year and plotted them over time in Fig. 4. We observe that for physics majors, these four constructs (Figs. 4a, 4b, 4c, and 4d) remain remarkably consistent throughout the physics majors' studies except for some slight declines over time in self-efficacy and identity. There are also some noticeable increases among graduate students over undergraduate students; however, note that graduate students comprise a different cohort than the undergraduate students, which could explain some of these differences. In particular, perceived recognition (Fig. 4b) and interest (Fig. 4c) show very little change even from the earliest time point (year 1 pre) to the latest (Grad post). For physics majors, perceived recognition ranges from a minimum of 3.23 in year 1 post to a maximum of

3.40 in year 4 pre, while interest ranges from a minimum of 3.45 in year 4 post to a maximum of 3.69 in year 1 post. Self-efficacy (Fig. 4a) shows a slight drop from year 1 pre (at 3.17) to year 1 post (3.00), but after that point it remains relatively consistent through Grad post, with slight decreases from each pre survey to the following post survey followed by a slight increase in the next year's pre survey. Overall, the maximum self-efficacy for physics majors occurs in year 1 pre at 3.17 and the minimum in year 3 post at 2.88. Students' physics identity (Fig. 4d), while still largely consistent over time, shows a slight decline from the maximum 3.63 in year 1 pre to the minimum 2.98 in year 4 post (i.e., the span of undergraduate years). This trend is reversed for graduate students whose mean physics identity score is slightly higher than the year 4 post responses of undergraduate students (keeping in mind that these cohorts are different students and a selection effect could explain these differences).

We now consider the constructs related to the physics learning environment, namely peer interaction (Fig. 4e) and students' sense of belonging (Fig. 4f). Both of these constructs are presented in a slightly different manner than the other constructs in Fig. 4. Peer interaction (Fig. 4e) is only included on the post surveys since the survey items pertain to students' experiences and interaction with their peers throughout the course. Student responses to peer interaction range from a minimum of 2.73 in year 2 post to a maximum of 3.05 in year 1 post. Moreover, it appears that the minimum reported peer interaction in year 2 post is somewhat misaligned with the rest of the responses, which are more consistently near 3 points on average. We will consider possible explanations for this after examining the other factor related to the perception of the learning environment, i.e., sense of belonging (in Fig. 4f, note that the scale is 1–5 for this factor). Figure 4f shows a quick decline from the maximum of 4.09 in year 1 pre to the minimum of 3.46 in year 2 post. However, this quickly returns to 3.96 in year 3 pre and remains relatively consistent from that point forward.

Both Figs. 4e and 4f point to the second year as being a somewhat more difficult time for the students' relationships with their fellow physics majors. One hypothesis for this drop in year 2 peer interaction is the dramatic change in the classroom population between year 1 (which is largely dominated by engineering students) and year 2 (when the majority of students are physics majors). Another possible reason is that in the introductory year 1 courses, students attended recitation sections in addition to lecture, wherein they were among a smaller group of students from the class. These recitation sections are not present in physics courses beyond the first year and so the year 2 courses may coincide with a transition period (which may even be marked by some students beginning to recognize the importance of establishing working groups among their physics major peers) while the students adjust to smaller class sizes in lecture instead of large lecture classes and smaller recitation sections. Interestingly, the remainder of the constructs in Fig. 4 shows that the other internal motivational characteristics of these students including self-efficacy, interest, and identity are not as susceptible to rapid change during the second year.

Next we turn to the first year responses to compare majors to non-majors and answer **RQ2**. The two sets of responses that appear in Fig. 4 for year 1 pre and year 1 post are from the same calculus-based physics courses, with those students who were identified as physics majors separated from the remaining non-physics majors. We find that on all measures, physics majors respond more positively on average than their non-physics major peers, though the difference between the two means is smaller in self-efficacy than in the other three constructs. In particular, for identity (Fig. 4d) the difference is 1.03 in pre and 0.85 in post; for perceived recognition (Fig. 4b) the difference is 0.70 in pre and 0.67 in post; and for interest (Fig. 4c) the difference is 0.67 in pre and 0.78 in post. However, majors and non-majors respond much more similarly in self-efficacy measures, with majors only 0.34 and 0.37 points higher on average in pre and post, respectively. Similarly, the responses to the peer interaction and sense of belonging questions are more similar between majors and nonmajors than the items relating to perceived recognition, interest, and physics identity. In particular, physics majors only respond on average 0.17 points higher than nonmajors on peer interaction in year 1 post, and on average 0.46 points higher in pre and 0.14 points higher in post on sense of belonging than nonmajors. One hypothesis for why this is the case is that even though physics is not their major, nonmajors in these calculus-based physics courses are the majority, with most students being engineering majors, and so their sense of belonging and perception of positive peer interaction should be bolstered by the presence of many of their peers from other common courses required by the engineering program.

Turning next to **RQ3** and investigating the predictive relations, we use multiple linear regression models to test the framework of physics identity predicted by interest, perceived recognition, and self-efficacy in accordance with previous studies of physics identity [116, 115, 142, 143, 145]. Table 6 shows six such models, one for physics majors in each year of study and one for non-majors in first-year physics courses. In every case, over 50% of the variance in physics identity is explained by the predictors (as indicated by R^2 in Table 6).

The findings for non-majors in year 1 shown in Table 6 reflect prior research for introductory physics including both majors and non-majors [116, 115, 142, 143, 145], and show that the primary predictor of physics identity is perceived recognition, while interest and self-efficacy are also statistically significant predictors (but with smaller regression coefficient β than perceived recognition). Focusing just on majors in year 1, we find a somewhat different model wherein perceived recognition is the sole predictor of physics identity. One hypothesis for this finding is that physics majors in their first year have more homogeneous levels of interest and self-efficacy in physics while there is more variation in their perceived recognition. We note, however, that there is also work to be done in improving student self-efficacy, as Fig. 4a shows that on average, a full point improvement is possible on the 1–4 Likert scale for the self-efficacy of physics majors.

Progressing to year 2, for physics majors (Table 6), perceived recognition remains the top predictor of identity while self-efficacy becomes a statistically significant predictor. Following this, for physics majors in years 3 and 4, interest becomes a statistically significant predictor alongside perceived recognition and self-efficacy, with the standardized β falling into a similar pattern as seen among non-majors in year 1. Finally, the model for graduate students shows that interest once again drops out as a significant predictor, leaving only perceived recognition and self-efficacy as the statistically significant predictors of physics identity for graduate physics students. One hypothesis for this is that, as in the case with first year undergraduates, first year graduate students have a more homogeneous level of interest in physics while their perceived recognition and self-efficacy vary from student to student.

Our main finding from these multiple regression models predicting identity is that perceived recognition is universally the top predictor of physics identity. This is the only predictor of identity for physics majors during the first-year, with later years approaching a regression model that resembles those found among a general population in introductory physics students in prior studies [116, 115, 142, 143, 145].

In order to answer **RQ4**, we further tested the relationships between self-efficacy, peer interaction, and sense of belonging in Tables 7, 8, and 9. Since these motivational constructs have complex inter-dependencies, we use these multiple regression models in order to understand how these relationships evolve over time.

Beginning with Table 7, we see that self-efficacy is primarily predicted by perceived recognition and sense of belonging, with peer interaction only being a significant predictor for non-majors and among physics majors in year 2. As with Table 6, every model reported in Table 7 explains more than 50% of the variance in self-efficacy. In years 1 and 2, recognition and sense of belonging predict the self-efficacy of physics majors with roughly similar standardized β. However, among year 3 and 4 physics majors and first-year graduate students, the best predictor of self-efficacy is sense of belonging, with the standardized regression coefficient β of perceived recognition slowly reducing over time, though always remaining statistically significant.

Next, we consider how peer interaction depends upon the other constructs in Table 8. We find that among physics majors, the only statistically significant predictor is sense of belonging, which explains between 33% and 42% of the variance in peer interaction for physics majors in every year except year 4, where only 16% of the variance in peer interaction is explained. Notably, neither perceived recognition nor self-efficacy were statistically significant predictors of peer interaction at the current level of statistical power.

Finally, in Table 9 we explore how sense of belonging depends on the other motivational characteristics. We find that in every model except for year 2 physics majors, the best predictor of sense of belonging is self-efficacy, with peer interaction being the next best predictor (except in year 4 when it is not statistically significant). The results for year 2 physics majors are once again an exception, as we saw when predicting physics self-efficacy in Table 7. One possible reason for these results in year 2 could be the lower sense of belonging in year 2 observed in Fig. 4.

3.4 Summary and Implications

In this cross-sectional study, there are two main effects that we observe when considering the average responses to the survey items (see Fig. 4). The first is that, by and large, physics majors from first-year undergraduate to first-year graduate students respond to these items pertaining to identity, perceived recognition, self-efficacy, and interest in physics very consistently throughout their education (Fig. 4), which answers **RQ1**. From the survey validation, we know that the students were correctly interpreting these items. Therefore, these consistent responses over time indicate that students are constantly adjusting their perspective when responding to these items. For example, first-year physics students responding to the prompt "if I study, I will do well on a physics test" responded based on their belief about their performance on physics tests in their introductory physics course, while fourth-year physics students in a quantum mechanics course responded based on their belief about their performance on quantum mechanics tests (rather than how well they believe they could perform on tests from their earlier courses).

It is interesting that these motivational constructs across the full physics curriculum using this same survey are so consistent (Fig. 4). However, there are two notable exceptions to this consistency. First, while physics identity shows an overall slight decline from year 1 pre to year 4 post, there are prominent drops from year 1 pre to year 1 post and year 4 pre to year 4 post. Both of these occur during extremely important moments in these students' careers, e.g., in year 1 these students are considering whether or not to continue their studies of physics, while in year 4 these students are deciding the next stage of their career after graduation. These sharp declines during crucial moments call for efforts to encourage physics majors that they can succeed in physics – at all times throughout the physics curriculum and especially around these transition periods.

The second major finding from Fig. 4, which answers **RQ2**, is the consistently higher responses of physics majors relative to their non-physics major peers in the same courses. However, this difference is notably smaller in self-efficacy, peer interaction (Figs. 4a, 4e; on the order of 0.3 points on the 1–4 Likert scale), and sense of belonging (Fig. 4f; on the order of 0.3 points on the 1–5 Likert scale) than in the other three constructs (Figs. 4b, 4c, 4d; on the order of 0.6–1.0 point on the 1–4 Likert scale). Many previous studies have emphasized self-efficacy as central to student success and retention, i.e., it is likely to influence not only student performance but also student retention and future career decisions [8, 9, 10, 11, 12, 13, 254, 253, 18, 145, 144, 177, 178, 214, 292, 24, 46, 157, 176, 272]. Our findings suggest that calls for physics departments to do more in order to boost student self-efficacy must be expanded to include all physics majors, rather than solely students in introductory physics courses.

With regard to **RQ3**, our findings from the multiple linear regression models in Table 6 show that perceived recognition is consistently the top predictor of physics identity, and this is especially true for physics majors in their first year. These findings expand and strengthen the findings of previous studies using the physics identity model [116, 115, 142, 143, 145] which called for physics instructors in introductory courses to make a concerted effort to recognize their students as people who can excel in physics. In particular, we find that this recognition is even more important for physics majors in the first year than non-majors and first-year graduate students, and for undergraduate students beyond the first year, it is just as important as it is for non-majors in the first year courses. Thus, we call upon all physics departments to implement measures to foster positive, encouraging interactions with all students to enhance their physics identity from introductory to graduate levels.

Further, in answering **RQ4**, we found in Tables 7, 8, and 9 that the relationships among self-efficacy, peer interaction, and sense of belonging were highly consistent over time. Future studies after more data are collected (and therefore statistical power increases) could explore these relationships further using statistical techniques such as structural equation modeling in order to include more complex relationships between the motivational constructs.

3.5 Acknowledgments

This research is supported by NSF Grant DUE-1524575 and the Alfred P. Sloan Foundation Grant G-2018-11183.

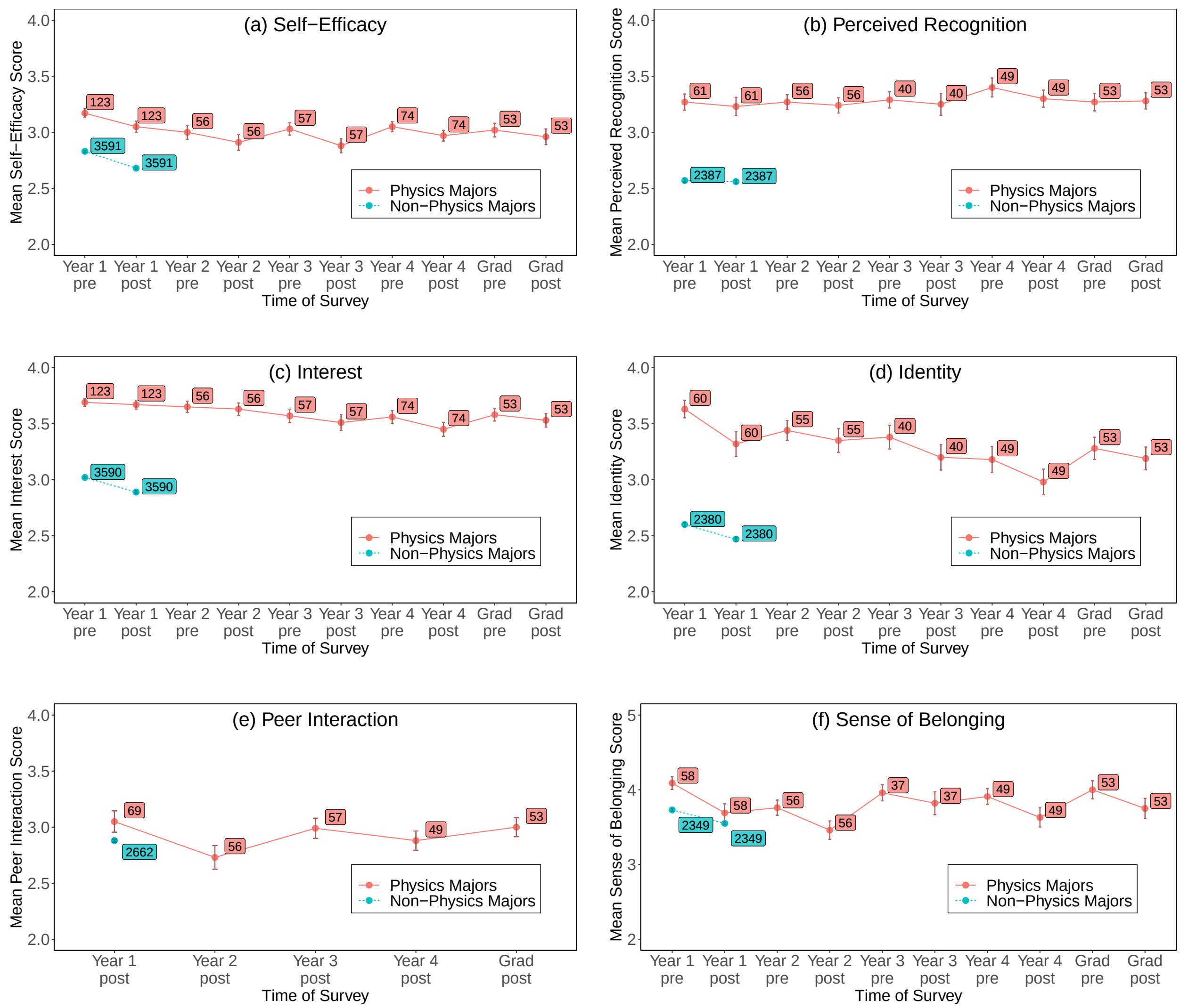


Figure 4: The mean scores of pre and post responses to (a) self-efficacy, (b) perceived recognition, (c) interest, (d) identity, (e) peer interaction, and (f) sense of belonging items are plotted along with their standard error for physics majors in each listed year of study. The responses are on a Likert scale of (a) through (e) 1–4 and (f) 1–5. Vertical axis ranges have been restricted for clarity. Since the questions pertaining to peer interaction were only given in the post survey, only post scores are plotted for those responses. The sample size is reported next to each point. Lines connecting points are drawn as guides to the eye.

Table 6: Summary of findings from multiple regression models predicting physics identity from interest, perceived recognition, and self-efficacy among physics majors in each year of study and non-physics majors in the first year. Each column shows the results from a regression using data for the specified year. Reported are the standardized regression coefficients (β) for each predictor along with the number of students (N) and fraction of variance explained in physics identity (R^2). The p-values are specified using asterisks, with non-significant predictors trimmed from the model and replaced by "–".

Standardized β predicting Physics Identity

Motivational Construct	Year 1 Non-majors	Year 1	Year 2	Year 3	Year 4	Grad
Interest	0.22^{***}	–	–	0.28^{**}	0.30^{***}	–
Recognition	0.50^{***}	0.73^{***}	0.45^{***}	0.43^{***}	0.39^{***}	0.62^{***}
Self-Efficacy	0.22^{***}	–	0.35^{**}	0.21^{*}	0.27^{**}	0.34^{***}
N	3287	94	65	71	71	58
R^2	0.65	0.53	0.52	0.60	0.57	0.74

$^{*}p < 0.05$ $^{**}p < 0.01$ $^{***}p < 0.001$

Table 7: Summary of findings from multiple regression models predicting physics self-efficacy from perceived recognition, peer interaction, and sense of belonging among physics majors in each year of study and non-physics majors in the first year. The conventions and notations match those of Table 6.

	Standardized β predicting Physics Self-Efficacy					
Motivational Construct	Year 1 Non-majors	Year 1	Year 2	Year 3	Year 4	Grad
Recognition	0.33^{***}	0.50^{***}	0.39^{***}	0.36^{***}	0.19^{*}	0.22^{*}
Peer Interaction	0.16^{***}	–	0.23^{*}	–	–	–
Sense of Belonging	0.45^{***}	0.45^{***}	0.30^{*}	0.54^{***}	0.64^{***}	0.63^{***}
N	3013	76	65	70	69	57
R^2	0.62	0.73	0.57	0.57	0.54	0.59

$^{*}p < 0.05$ $^{**}p < 0.01$ $^{***}p < 0.001$

Table 8: Summary of findings from multiple regression models predicting peer interaction in physics from perceived recognition, self-efficacy, and sense of belonging among physics majors in each year of study and non-physics majors in the first year. The conventions and notations match those of Table 6.

	Standardized β predicting Peer Interaction					
Motivational Construct	Year 1 Non-majors	Year 1	Year 2	Year 3	Year 4	Grad
Recognition	0.10^{***}	–	–	–	–	–
Self-Efficacy	0.27^{***}	–	–	–	–	–
Sense of Belonging	0.32^{***}	0.61^{***}	0.65^{***}	0.57^{***}	0.41^{***}	0.59^{***}
N	3013	76	65	70	69	57
R^2	0.37	0.37	0.42	0.33	0.16	0.35

$^{*}p < 0.05$ $^{**}p < 0.01$ $^{***}p < 0.001$

Table 9: Summary of findings from multiple regression models predicting sense of belonging in physics from perceived recognition, self-efficacy, and peer interaction among physics majors in each year of study and non-physics majors in the first year. The conventions and notations match those of Table 6.

	Standardized β predicting Sense of Belonging					
Motivational Construct	Year 1 Non-majors	Year 1	Year 2	Year 3	Year 4	Grad
Recognition	0.11***	–	0.29**	–	–	–
Self-Efficacy	0.52***	0.60***	0.26*	0.53***	0.71***	0.60***
Peer Interaction	0.22***	0.28***	0.43***	0.34***	–	0.27**
N	3013	76	65	70	69	57
R^2	0.65	0.53	0.55	0.60	0.51	0.61

$^{*}p < 0.05$ $^{**}p < 0.01$ $^{***}p < 0.001$

4.0 Engineering Students' Performance in Foundational Courses as a Predictor of Future Academic Success

4.1 Introduction

4.1.1 Background and Theoretical Framework

Relationships among courses are critical in the design of a curriculum, especially for interdisciplinary fields such as engineering that integrate many areas of science and mathematics. Further, there can be a struggle to fit all requisite foundational courses into earlier coursework (whether in upper secondary or early university) while also allowing for interest-based exploration. Knowledge of strong relationships between foundational courses and later coursework can support overall curriculum revision efforts as well as personalized learning decisions. Furthermore, it is important to quantitatively investigate how performance in these courses correlates with later performance in the curriculum in order to confirm the assumptions that these courses are foundational to a successful education in engineering. Evaluation of the strength of course relationships in an existing curriculum can be accomplished using educational data mining of large institutional data and learning analytics. In this paper, we present a methodology for implementing such learning analytics with existing institutional data and apply the methodology to the data from one US-based institution to test foundational questions about the nature of predictive relation between first year science and mathematics and later engineering courses. While the study is implemented within one US-based engineering program, the broader methodology described can be applied to curricula at any institution.

Engineering schools are increasingly recognizing the importance of evidence-based approaches to improve student learning and ensure that all students have the opportunity to excel regardless of their background [37, 38, 39, 138, 247, 255]. In addition, preparing a diverse group of engineers who are able to embrace the challenges and opportunities of the 21st century is essential to strengthening the Science, Technology, Engineering, and Mathematics (STEM) workforce. Meeting these goals requires that institutions take a careful look at the extent to which engineering education is equitable and inclusive and provides adequate support, advising and mentoring to all students from diverse backgrounds who have traditionally been left out to ensure that even those students who had less than ideal opportunities at the K-12 level have the opportunity to excel in engineering undergraduate programs.

Making appropriate changes to curricula based upon data requires holistic consideration of how an undergraduate engineering school is currently succeeding, including how prior foundational courses predict future performance in engineering courses. Such information can be useful regardless of the theory of change an engineering school adopts and implements. At the same time, with advances in digital technology in the past decade, data analytics can provide valuable information that can be useful in transforming learning for all students [6, 217].

Much prior research on engineering education has focused on: 1) evidence-based classroom practices to improve various facets of learning [163], 2) how to evaluate the effectiveness of different pedagogical approaches [89, 196], and 3) the balance of teaching theory and practice [256]. But there has been much less focus on how student performance in different subsequent courses builds upon prior courses. Information obtained from data analytics on large institutional data in these areas can be an important component of understanding, e.g., the role foundational courses play in later engineering performance as well as contemplating strategies for strengthening these ties, and improving student advising, mentoring, and support.

Studies of curricula as a whole have focused on more general outcomes such as enrollment, degree attainment, and retention rates within engineering programs [21, 92, 94, 112, 189, 202, 203]. Of particular relevance is the extensive research conducted using the Multiple-Institution Database for Investigating Engineering Longitudinal Development (MIDFIELD) [165, 167, 211]. The MIDFIELD dataset is a rich source of curricular data that allows for multi-institution investigations into how students across engineering fields and from different demographics (e.g., women and underrepresented racial and ethnic minorities) thrive throughout a curriculum and the relative frequency at which they complete their degrees and obtain engineering jobs [44, 45, 166, 170, 192, 210, 212, 222].

A yet under-explored question is how well do foundational engineering, science, and mathematics courses (which are sometimes considered weed-out courses) predict performance of students in later engineering courses. Because of the variation across institutions in foundational course requirements [181], this kind of investigation naturally begins with an investigation of a particular engineering school's first-year curriculum. Such an investigation can be useful for other institutions that may perform similar analyses (but adapted to their required foundational curricula) in order to contemplate strategies for improving engineering education in a holistic manner. We hypothesize that some lessons may prove to be quite general since there is considerable overlap in foundational courses. For example, early required coursework in physics and mathematics is very similar in most countries, although these may be completed during upper secondary rather than first year of the engineering degree in some countries. Within most large US engineering programs, the more specific structure of required foundational courses is more consistent. These foundational courses are required under the assumption that the later engineering courses would build on these subjects. Below, we summarize other relevant prior research literature before laying out our research questions.

4.1.2 Relevant Prior Literature Review

Very few studies to our knowledge have focused on the relationships between specific science and mathematics courses and subsequent engineering courses using large data analytics. Those few studies have focused on aspects of course relationships such as, in the case of Reeping, Knight, Grohs, and Case [227], enrollment patterns and grades earned by students repeating courses. One analogous study focused on science course outcomes and explored the relationship between foundational science and mathematics courses taken in high school and performance in introductory college science courses in biology, chemistry, and physics [235]. These researchers used a linear regression analysis with the number of years of high school instruction in each of these subjects as well as mathematics as the independent variables and course performance in introductory college science courses as the dependent variables. They found that the years of instruction in each subject predicted student college success in that same subject, but the only cross-disciplinary correlation was the years of instruction in mathematics predicting performance in every introductory college science course. However, it is possible that these introductory college science courses were taught in a way that made few assumptions about past learning. A different result may emerge for the relationship between foundational courses and engineering courses.

From the student perception side, a number of studies have noted that engineering students often question the relevance of foundational courses to the work of engineers [96, 289]. In particular, engineering students display mixed perceptions of the importance of mathematics to their studies in engineering [67, 281], while engineering faculty members, who perceive various topics in mathematics to be important to an education in engineering, perceive that engineering students graduate with insufficient competency in these topics [113]. Further, this doubt of relevance of foundational science and mathematics courses induces motivational problems for engineering students that then lead to poor course performance and attrition [21]. Having empirical evidence for the relevance of foundational courses to later engineering coursework using large institutional data and subsequently conveying these findings to the students has the potential to improve student motivation.

There can be many reasons why there may not be a transfer of knowledge acquired in one course to improved performance in a later course that is theoretically related [79]. First, the actual overlap in terms of specific knowledge and skills can be minimal (e.g., the main area of overlap may be the last part of the foundational course which could be treated as an optional special topic). Second, there can be superficial differences in the way key concepts and skills are discussed or represented (e.g., notational differences for how derivatives or energy are represented) that prevent students from discerning connections between courses. There is a wealth of literature in the cognitive and learning sciences showing that students often fail to spontaneously transfer relevant conceptual knowledge and skills to novel problems with new surface features or those presented in different contexts [16, 101, 233, 75]. Third, the instruction in either the foundational course or the later courses may not frame the knowledge and skills that are being learned in general ways that lead students to be more likely to generalize and make connections [91]. Thus, there may be an empirical foundation to students' belief that foundational coursework is not actually helpful for later engineering coursework.

On a related matter, there is the question of the form of the relationship between courses, which is important for effective model building but also has practical consequences. The simplest possible forms of course relationships are threshold and linear functions. Many universities require a minimum letter grade of C in order to move on in a sequence (or a mid-level exam outcome in the case of advanced coursework taken in secondary, such as Advanced Placement courses in the US). Informally, students sometimes provide practical advice to each other that only a C is even needed in order to succeed in later coursework ("Cs get degrees") [155, 154]. Each of these is suggestive of a possible threshold function in which higher performance above a certain level (e.g., an A rather than a C or B) does not translate into higher performance in later courses. Alternatively, the relationship among courses may be linear. Huang and Fang [130] compared four different predictive models of academic performance with multiple linear regression as the simplest model and found no significant advantage of the other models over multiple linear regression for predicting the academic performance of large groups of students.

A central methodological issue in studying relationships among courses is controlling for the many correlated sources of performance. In addition to examining in parallel the many foundational courses that could influence a given target course (e.g., physics, chemistry, calculus, and a MATLAB foundational course for a target introductory Materials Engineering course), there are also more general student factors such as general intelligence, mathematical skill, overall study skills, and general academic motivation. In particular, performance in any one course is expected to be correlated with performance in any other course simply by virtue of these more general factors (e.g., students with high academic motivation may perform at higher levels in most courses). Such general factors can be modeled and controlled for in regressions using indictors of general knowledge, motivational, and general academic skills such as high school GPA, performance on the SATs, and cumulative university GPA. For example, Huang and Fang [130] found that the best predictor for a second-year engineering dynamics course was the students' cumulative GPAs, although they also found that grades in prerequisite courses mattered above and beyond cumulative GPA.

4.1.3 Goals and Research Questions

The current effort develops and tests an approach to using basic statistical techniques on large longitudinal institutional datasets that are increasingly being made available for researchers and practitioners to address problems of both theoretical and practical consequence. Our research focuses on building predictive models for a wide range of foundational engineering curriculum courses in each semester through the first two years using the data from one mid-sized, research-intensive university in the US. Furthermore, we investigate these predictive relationships for selected courses from the two largest engineering departments (or disciplines), namely Mechanical Engineering and Materials Science (MEMS) and Electrical and Computer Engineering (ECE). These courses are typically thought to rely heavily on mathematics and physics (i.e., are especially likely to have concerns about poor student performance in later coursework because of earlier weak performance). Understanding the hierarchy of predictive relationships from foundational courses to subsequent engineering courses provides a useful context for curricular evaluation and advising students as they progress through the first two years of instruction in engineering.

4.1.4 Research Questions

For the two engineering disciplines we focused upon, our research questions (RQs) to guide the investigation are:

RQ1. Do foundational course grades relate via a linear, threshold, or some other function to grades in later engineering curriculum courses?

RQ2. To what extent does student performance in foundational courses predict performance in second-year engineering courses above and beyond general student performance factors?

RQ3. Which foundational courses are most important to course achievement in core second-year engineering courses?

In addition to individually addressing important theoretical and practical questions, as a sequence, these RQs also address the systematic development of appropriate statistical models, moving from working out appropriate functional forms to testing larger patterns of course relationships.

4.2 Methodology

4.2.1 Participants

Using the Carnegie Classification of Institutions of Higher Education [133], the university at which this study was conducted is a public, high-research doctoral university, with balanced arts and sciences and professional schools, and a large, primarily residential undergraduate population that is full-time, more selective, and low transfer-in. Deidentified application information and course grades data were provided by the university on all first-year engineering students who had enrolled from Fall 2009 through Spring 2018. This data provision is part of a larger NSF-funded inter-departmental effort towards improving education at the university, and the form of the data is similar to the MIDFIELD dataset. The full sample for the current study consists of 5,348 engineering students, identified by having taken either of the introductory engineering courses. The subset of this sample for analyses into the second-year curriculum of two engineering departments consisted of 2,825 students, which includes all students who took at least one third-semester engineering course listed in Table 10 in addition to the introductory engineering courses. The full sample of students includes 27% female students and had the following race/ethnicities: 80% White, 8% Asian, 5% African American, 2% Latinx, and 5% Other. The mean age at the beginning of the student's first year was 18.9 years (SD = 1.7 years), reflecting a population of students who were predominantly attending college immediately after completing high school.

4.2.2 Curricular Context

This study is primarily focused on the engineering curriculum of the Department of Mechanical Engineering and Materials Science (MEMS), one of the largest engineering departments, to produce statistically robust results. One course is also considered from the

Table 10: The relevant portions of the MEMS and ECE curricula along with shortened names used in other tables and figures and total sample size (N) for each course. Bolded are the target second-year engineering courses. Semesters 3 and 4 are further divided into MEMS and ECE.

Semester	Full Course Name	Short Name	N
1	Basic Physics for Science and Engineering 1	Phys 1	5003
	Introduction to Engineering Analysis	Engr 1	5348
	Analytic Geometry and Calculus 1	Calc 1	4214
	General Chemistry for Engineers 1	Chem 1	4767
2	Basic Physics for Science and Engineering 2	Phys 2	4069
	Introduction to Engineering Computing	Engr 2	4435
	Analytic Geometry and Calculus 2	Calc 2	4009
	General Chemistry for Engineers 2	Chem 2	3706
3 (MEMS)	Analytic Geometry and Calculus 3	Calc 3	3703
	Introduction to Matrices and Linear Algebra	Lin Alg	2572
	Materials Structure and Properties	Mat Structure	1582
	Statics and Mechanics of Materials 1	Mech 1	1995
3 (ECE)	Linear Circuits and Systems 1	ECE Circuits	803
3 (ECE) / 4 (MEMS)	Differential Equations	Diff Eq	3607
4 (MEMS)	Statics and Mechanics of Materials 2	Mech 2	851
	Electrical Circuits	MEMS Circuits	775

Department of Electrical and Computer Engineering (ECE) because the department offers a parallel version of a course on circuits in a different order relative to mathematics courses than the MEMS circuits course; a comparison of performance across these courses enables a robust test of foundational course presence/absence that is not confounded with selection artifacts (e.g., the students that delay foundational courses typically have other challenges). The curriculum analyses focus on year 1 (foundational courses, including two engineering courses designed to teach students MATLAB, C++, and Python) and year 2 (additional foundational courses in mathematics and core discipline-specific engineering), where the connections to foundational courses are potentially most robust and students are most likely to struggle.

Table 10 shows the relevant portions of these curricula, including the first-year courses taken by all engineering students and the selected second-year courses in engineering and mathematics (the only foundational topic that extends beyond the first year in MEMS). Table 10 also shows the number of students within the sample that took each of the courses whose predictive relations are analyzed in this study. In principle, all of the first-year courses are required for all engineering students; however, in practice this is not always the case. A student may lack a grade in the data for many reasons, ranging from skipping the course with Advanced Placement (AP) credit or taking the course at another university. Since not all of these students will have taken the remaining courses used in the study, Table 10 provides an upper bound on the N for statistical tests used because we employ list-wise deletion for our analysis, which will drop a student from a given analysis if the student is missing any one of the variables. Some courses, such as Statics and Mechanics of Materials 1, appear in multiple engineering curricula which may have different requirements for second-year mathematics courses; the sample size (N) for statistical analyses involving these courses will be substantially reduced depending upon which courses are included as predictors since each participant in a given model is required to have data for all variables.

4.2.3 Measures

4.2.3.1 High School Academic Achievement From the information submitted by the students as part of their applications to the university, the university provided several key pieces of data that are commonly used by universities in the US in admission decisions (because they are predictive of success in university coursework). High school grade point average (HS GPA) is the mean grade across all high school courses (grades 9-12) on a weighted 0-5 scale, which involves adjustments to the base 0-4 scale for AP and International Baccalaureate courses. These adjustments were performed prior to our acquisition of the data. In addition, students take one of two different standardized assessments: the Scholastic Achievement Test (SAT) or the ACT. There are multiple components to each, but we focused on the mathematics achievement component, and converted ACT scores (1-36) into SAT scores (200-800). For the sample of engineering students included in this study, the mean SAT Math score was 690 (SD = 60), and the mean HS GPA was 3.99 (SD = 0.41).

4.2.3.2 Grades The primary data provided are grade points (GPs) earned in all courses at the university, the semester and specific class section in which each course was taken, and the grade point distributions (mean, sample size, and standard deviation) for each class (used to remove effects of instructor grading variation). GPs are on a 0-4 scale (F = 0, D = 1, C = 2, B = 3, A = 4) where the suffixes '+' and '–' add/subtract 0.25 (e.g., B+ = 3.25) except A+, which has a GP of 4. In order to reduce effects from the particular year and instructor of each course, grade data were "z-scored" using the GP earned by the student along with the mean μ and standard deviation σ of the student's specific class section to calculate $z = (\text{GP} - \mu)/\sigma$. The z-score is in units of standard deviation. Though the university does allow students to retake courses and replace earlier grades, the data provided to us are the initial raw grade data.

In addition, each student's cumulative STEM GPA was calculated for each semester, defined as the mean GP on all science, engineering, and mathematics courses taken at the university up to that point in time. This measure is meant to capture overall STEM performance as a mixture of general problem solving and reasoning skills, study skills, general motivation in STEM, and engineering-specific competencies. Including such a measure allows the specific value of knowledge and skills derived from foundational courses to be separated from general skills and motivation.

4.2.4 Analysis

4.2.4.1 Inclusion Criteria Engineering curricula have a large number of required courses and relatively little flexibility in the order in which courses are taken due to complex sequences of course prerequisites. Nonetheless, students sometimes take courses earlier than expected (e.g., due to skipping an earlier course based upon advanced high school coursework) or later than expected (e.g., due to failing a core course or taking a reduced course load). The core chronology of the engineering curriculum was used as a filter on included data both to maintain the directionality of the predictive relationships and to inform which attempt should be used if a student attempts a course multiple times: each "target course" that appears includes only students' first attempt of that course, and the semester in which that first attempt occurred was used to select the latest attempt of each "predictor course" that occurred prior to or concurrent with the target course. This chronological enforcement was used in every analysis in this study, all of which contain one target course with one or more predictor variables. Note that the cumulative STEM GPA through one semester prior to the target course was also used as a predictor to control for the general student skills and motivation; we did not use STEM GPA for the whole university degree because predictors generally should not include performances that happen later in time than the dependent variable.

4.2.4.2 Qualitative Model Building To get a qualitative sense of the relationships between courses in order to select appropriate statistical models, we generated histograms which first bin students by their letter grade in a predictor course, then within each of those bins, the students are further binned by their letter grade in a target course. Results show no qualitative difference if the mean grade points used are z-scored, so the unaltered grade points are used to improve readability. Furthermore, in order to keep the number of bins manageable in this analysis, we grouped together letter grades that differ by a plus or minus sign. For example, the grades B-, B, and B+ were all grouped together as B for some analyses.

4.2.4.3 Selecting the Appropriate Regression Function In order to determine if the nature of these relationships was a linear, threshold, or some other function (**RQ1**), we also generated graphs plotting the (continuous scale) mean grade points earned in a target course against (discrete) grades earned in various prior or concurrent courses that may predict the target course. These graphs are shown both with and without the +/− distinctions, for example, combining B-, B, and B+ letter grades in predictor courses.

4.2.4.4 Model Building Procedure The regression analyses were performed using the "regress" function in Stata version 15.1 [111]. In order to address **RQ2** and **RQ3**, we sought best fitting multiple regression models [195, 201] with the (z-scored) grades of the target course as the dependent variable and (z-scored) grades of predictor courses along with general academic performance variables (cumulative STEM GPA, high school GPA, and SAT Math scores) as the independent variables. Note that the list of predictor courses includes all courses taken prior to the target course as well as concurrent mathematics courses. Concurrent mathematics courses have been included since research has suggested there are synergistic benefits to taking concurrent courses utilizing similar content [132]. For target courses, we used the first attempt of the course as the dependent variable, while for predictor courses we used the latest attempt of the course prior to or concurrent with the first attempt of the target course. Initially, all courses prior to or concurrent with the target course were included in the model, and then a modified stepwise-deletion procedure was applied: 1) at each step, the independent variable with the largest $p > 0.05$ was dropped; 2) the regression was re-run with the remaining independent variables; 3) this was done iteratively until all of the remaining independent variables had $p < 0.05$; and 4), because automatic stepwise procedures for model building can get stuck in local minima, each dropped independent variable was put back in the model along with the pared-down list of remaining independent variables in order to ensure that the regression coefficient was not statistically significant (i.e., the previously dropped independent variables' still had $p > 0.05$ when individually reinserted).

The model building procedure was conducted twice for each target course: once without cumulative STEM GPA and once again with it included. Cumulative STEM GPA is meant to control for a general, overall student performance dimension, addressing a combination of study skills, academic motivation, and outside competing pressures (e.g., part-time jobs, hobbies, and family obligations). However, since it is derived from the grades in the predictor courses, it partially also captures knowledge and skills from each foundational course. Thus, including it as a control variable may control too much, potentially producing underestimates of the effects of the foundational courses. Therefore, models with and without this predictor are presented to provide upper and lower bounds on the effects of each foundational course.

From these regression analyses, we report the list of β coefficients (standardized to reflect connection strength in effect size units) for significant predictors along with their p-values as well as the number of students remaining in each regression, N, and the adjusted proportion of variance in the target course explained by the regression, R^2_{adj} [195, 201]. In all models, R^2_{adj} falls between 0.35 and 0.55 and is less than 0.01 smaller than the non-adjusted R^2. Since the best fitting models may include general academic variables, **RQ2** is addressed by examining the combined standardized beta coefficients of the course predictors. Which course predictors remain significant and their relative strengths as indicated by standardized beta coefficients address **RQ3**.

4.3 Results and Discussion

4.3.1 General Predictive Relationships Between Courses

First, leaving aside for now issues of relative predictive power and controls, every pairing of foundation and target courses showed a positive relationship in that higher foundational course grades were always associated with higher target course grades, with unimodal distributions. There were also interesting variations in the grade relationship distributions found in foundational course-target course pairings. Figure 5 presents exemplars of the four qualitatively different variations. The differences in these histograms are most readily explained by the standard deviations of grades earned in the target and predictor courses. Figure 5a shows the target course Materials Structure predicted by Engineering 1, which has a particularly narrow and high grade distribution, leading to an unusually small but highly at-risk population in the C bin. On the other hand, Fig. 5b shows what happens when the target course, in this case Mechanics 1, has a narrow and high grade distribution: As and Bs are generally likely and it is primarily the ratio of As and Bs that is predicted by the predictor course. Finally, Fig. 5c and Fig. 5d show histograms with more typical grade distributions in both target and predictor, whether the predictor course has a low mean (Fig. 5c) or a high mean (Fig. 5d), indicated by the relative number of As, Bs, and Cs in the predictor course.

These kinds of figures may provide information that is useful for advising, especially Fig. 5d in which over half of the students who received a C in Linear Algebra (term 3 in the MEMS curriculum) went on to earn a C or worse in MEMS Circuits (term 4 in the MEMS curriculum). There are a few courses for which a C in the foundational course, although technically sufficient to go on to later courses that have it as a prerequisite, provides a strong predictive signal that the student is likely to struggle in that target course and therefore will likely need extra assistance. Academic advisors can also use this kind of information with their advisees to help motivate students to take the prerequisite courses more seriously. That being said, it is important to be careful to always use such data constructively, and in particular to avoid profiling students and contributing to low student competency beliefs, which may be the underlying problem.

4.3.2 Testing Linear or Threshold Functions

To better understand the nature of these relationships and answer **RQ1**, we examined line graphs for all foundation-target pairings, as is shown for Mechanics 1 as a target course in Fig. 6. We binned students by their letter grade in the predictor course, as was done for the histograms in Fig. 5, but then plotted the mean grade points earned by members of each bin in the target course rather than the full distribution. We observe from these line graphs

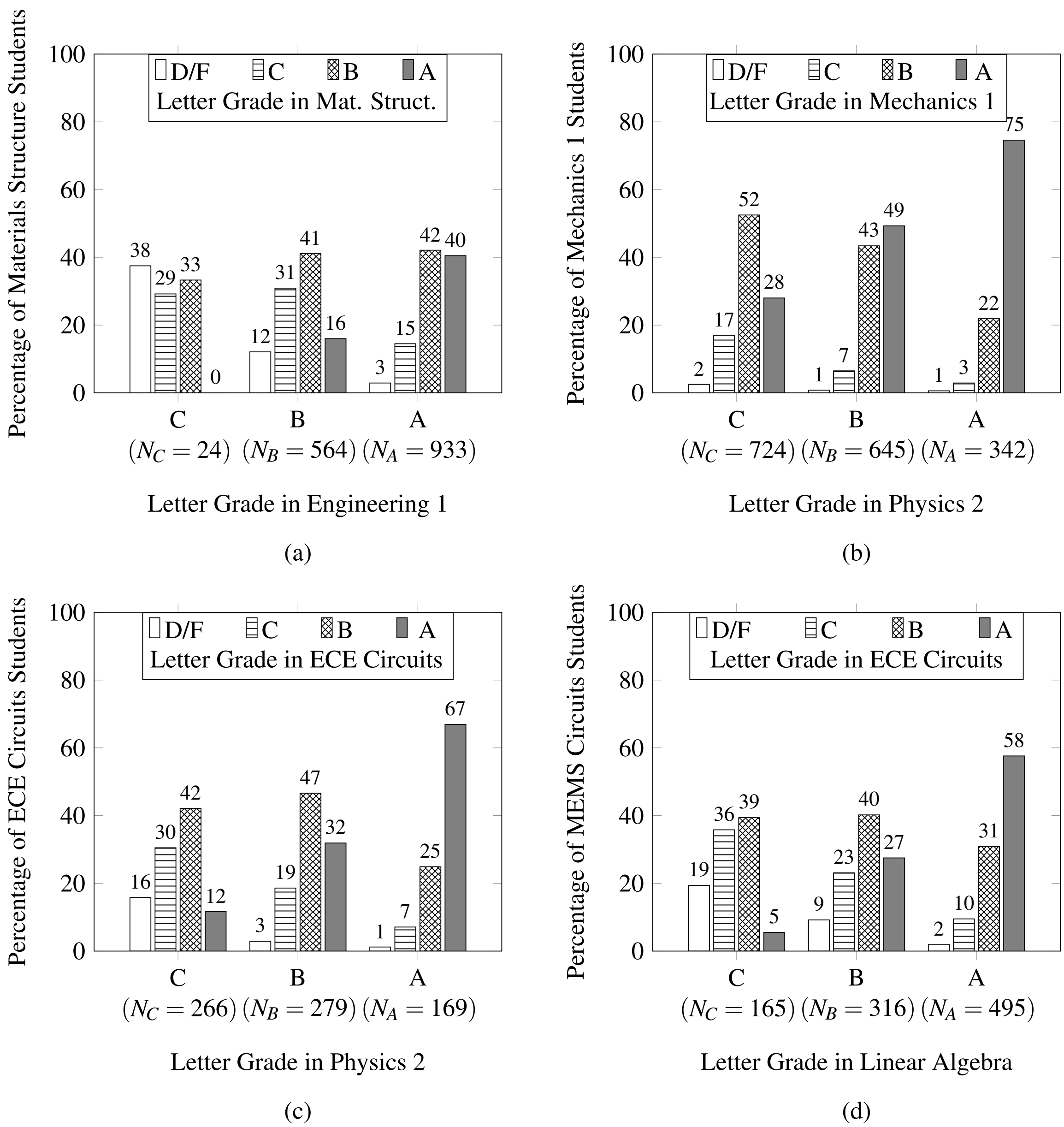


Figure 5: Exemplars of the different relationships between grades earned in pairings of predictor and target courses. Students are binned by their letter grade in the predictor course (horizontal) with the total number of students in each of these three groups (N_A, N_B, and N_C) shown below the letter. The percentage of each group that went on to earn an A, B, C, or D/F in the target course (vertical) is shown in the group of four bars above their grade in the predictor course. Each subfigure shows a different predictor and target pair indicated on the axis labels.

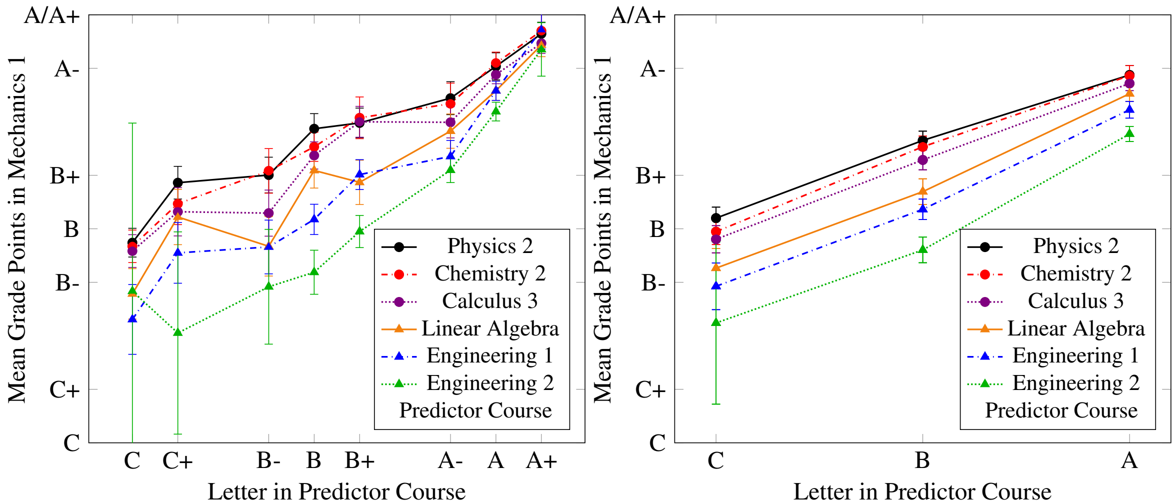
that the relationships are linear, which is especially clear when the bins are widened by removing plus and minus signs from letter grades to reduce measurement error; there were often large error bars associated with rarely given +/- grades). In every predictor-target course pairing, the relationships displayed similarly clear linearity. Note that the widened grade bins in Fig. 5 and Fig. 6b were used solely for this diagnostic purpose—the linear regression analyses used the more fine-grained grade data.

4.3.3 Model Building Results

Knowing that the courses are related linearly, multiple regression models were built using linear functions with the list of predictors consisting of high school GPA, SAT Math score, cumulative STEM GPA through the previous term, each prior course in the curriculum, and concurrent second-year mathematics courses. The results of the final predictive models for each target course were examined in terms of the standardized beta coefficients (β) to answer **RQ2** and **RQ3**. These β coefficients are standardized to range from -1 to 1, though all of the β coefficients we find in this study are positive; the values are similar to Pearson correlation values [195, 201]. Table 11 reports the regression results for each of the target second-year engineering courses in this study. In addition to the β coefficients, Table 11 reports the p-value of each β coefficient as well as the N and R^2_{adj} for each regression.

These results are reported both with and without the STEM GPA term included. Comparing these two sets of results, it is clear that chemistry, present only in the regressions without STEM GPA, was acting as a proxy for general factors (e.g., student ability or general academic motivation) while physics, mathematics, and engineering courses remain as predictors in both versions. This asymmetric reduction of β coefficients with the inclusion of a measure of general skills shows that these results are measuring the effects of these courses (in physics, mathematics, and engineering) beyond general skills and other general factors that influence student performance.

There is a natural categorization of the results of Table 11 dependent on the semester in which the target course is taken. For the semester 3 courses (ECE Circuits, Materials Structure, and Mechanics 1), the hierarchy of predictors is second-year mathematics first, followed



(a) (b)

Figure 6: Examining the nature of course relationships (linear, threshold, etc.) in order to inform statistical models. Students are binned by their letter grade in the predictor course (horizontal), then the mean grade points earned in the target course by the students in each bin are plotted vertically along with the standard error. The spacing of letter grades corresponds to the university's grade point values. This linear trend holds for every target/regressor pair in our analysis. Subfigure (a) shows the students binned by all letter grades, while subfigure (b) groups all students who earned, for example, C-, C, and C+ into a single C group.

Table 11: Results from the multiple linear regression analyses predicting selected second-year engineering courses. The numbers reported are the standardized β coefficients and p-values ($^{*}p < 0.05$,$^{**}p < 0.01$,$^{***}p < 0.001$). Each target second-year engineering course corresponds to two columns, with regressions run both with (a) and without (b) cumulative STEM GPA as an independent variable. "N/A" entries were excluded in the regressions because they occurred later in the curriculum; "–" entries were included but not statistically significant. Variance Inflation Factors (VIFs) are reported for the STEM GPA predictor as well as the maximum for all other predictors.

	Outcome Courses									
	ECE Circuits		Mat Structure		Mechanics 1		Mechanics 2		MEMS Circuits	
Regressors	*a*	*b*	*a*	*b*	*a*	*b*	*a*	*b*	*a*	*b*
Phys 2	0.15^{***}	0.12^{***}	0.16^{***}	0.13^{***}	0.10^{**}	0.10^{***}	0.08^{*}	–	0.10^{**}	–
Engr 1	0.17^{***}	0.09^{*}	0.11^{**}	0.07^{*}	0.09^{*}	–	–	–	–	–
Engr 2	–	–	–	–	0.13^{**}	0.11^{***}	–	–	–	–
Chem 1	–	–	–	–	–	–	0.13^{***}	–	0.10^{**}	–
Chem 2	0.12^{**}	–	0.13^{**}	–	0.11^{**}	–	–	–	–	–
Calc 3	N/A	N/A	0.18^{***}	0.16^{***}	0.21^{***}	0.14^{***}	–	–	–	–
Linear Alg	N/A	N/A	0.26^{***}	0.25^{***}	0.23^{***}	0.23^{***}	0.17^{***}	0.13^{***}	0.20^{***}	0.15^{***}
Diff Eq	0.41^{***}	0.35^{***}	N/A	N/A	N/A	N/A	0.38^{***}	0.27^{***}	0.35^{***}	0.27^{***}
Mech 1	N/A	N/A	N/A	N/A	N/A	N/A	0.15^{***}	0.09^{**}	N/A	N/A
STEM GPA	N/A	0.25^{***}	N/A	0.18^{***}	N/A	0.19^{***}	N/A	0.35^{***}	N/A	0.37^{***}
N	495	634	650	836	817	1063	599	735	632	706
R^2_{adj}	0.42	0.44	0.38	0.40	0.39	0.38	0.50	0.53	0.43	0.46
Max VIF	1.70	1.94	1.67	1.97	1.79	1.84	1.75	1.97	1.80	1.88
VIF (STEM GPA)	N/A	3.47	N/A	4.00	N/A	3.16	N/A	3.08	N/A	2.33

by STEM GPA, followed by a second second-year mathematics course if taken, followed by introductory engineering and physics courses. For the semester 4 courses (Mechanics 2, MEMS Circuits), STEM GPA has risen to be the top predictor followed by two second-year mathematics courses, then in the case of Mechanics 2 the last regressor is Mechanics 1, the prior course in the sequence. These trends indicate that the transfer of specific skills, at least in terms of what can be measured by this type of model, appears to be mostly limited to transfer from one semester to the next but not beyond, with the exception of Engineering 1, which occasionally will predict courses in semester 3 instead of Engineering 2. We note that the content of the introductory engineering sequence is not standardized across institutions, even within the US. At the studied institution, this sequence teaches students computer programming skills in an engineering context.

The magnitudes of the β coefficients in Table 11 range from low to medium, and correspondingly the variance explained in the target courses (i.e., R^2_{adj}), ranges from 38% to 53%. This is consistent with the findings shown in Fig. 5, where despite clear trends toward students maintaining similar grades from one course to another, there was some mobility for students. The variance not accounted for by the model may be partially due to the course-grained nature of the grade data itself. In addition, there should also be effects of pedagogical differences between instructors, TA quality, and life events. Still, despite their low to medium size, the majority of the coefficients estimated were highly statistically significant.

In order to test for multicollinearity problems, we followed up these analyses by calculating the Variance Inflation Factors (VIFs) of each regressor, which measure the degree to which the associated regression coefficient variances have been inflated due to collinearity with other independent variables [195, 201]. In every case, the VIFs of all regressors except those for cumulative STEM GPA are below 2.00, including concurrent mathematics courses, while cumulative STEM GPA ranges from 2.33 to 4.00. These values fall well below the commonly cited cutoffs of 10 [201] or the more recent recommendation of 5 [195]. Thus, the data were sufficiently large to overcome the challenges of estimating separable effects of correlated variables.

These multiple regression analyses were done not only with the core engineering courses as target courses, but also for each of the required second-year mathematics courses in the curriculum that could be foundational to other core engineering courses. Using all of these results together, Fig. 7 shows a visual representation of the curriculum with all statistically significant predictive relationships (from the regressions including cumulative STEM GPA) displayed as lines connecting the nodes representing the variables. Chemistry courses are not included in Fig. 7 because they were never statistically significant in the final models once STEM GPA was included; it may be that chemistry is more relevant for core courses in chemistry-oriented engineering majors (i.e., chemical, petroleum, or environmental engineering) in contrast to the core courses in physics-oriented engineering majors that were examined in this study.

A few basic patterns are observable within and between diagrams in Fig. 7. STEM GPA is a large predictor of all second-year courses, potentially because it becomes a more refined general-performance estimate with more courses contributing to it. The SAT Math variable drops out as a direct predictor after the first semester, later acting indirectly "through" STEM GPA. In contrast, high school GPA continues to have a small direct relationship to calculus courses and a large indirect effect through STEM GPA. Similarly, first-year courses can have not only direct effects on target second-year engineering courses, but also indirect effects mediated through the predictor course's effect on some other course that predicts the target. For example, in Fig. 7b we do not see a direct effect of Physics 2 on Mechanics 2, but it does have an indirect effect since Physics 2 predicts performance in Mechanics 1, which predicts performance in Mechanics 2. Finally, with only one exception, there are always direct connections within course sequences (e.g., Physics 1 to Physics 2; Engineering 1 to Engineering 2; Calculus 1 to Calculus 2; Mechanics 1 to Mechanics 2). The one exception is that Calculus 3 is predicted directly by neither Calculus 1 nor 2; here the inclusion of STEM GPA may have produced too conservative an estimate because it includes both Calculus 1 and Calculus 2 grades. However, Differential Equations maintains a connection to Calculus 2, so it may be that some other explanation is required (e.g., the more difficult multi-variate aspects of Calculus 3 are not based on remembering aspects of Calculus 1 and 2).

Addressing **RQ2**, course performance in every second-year mathematics course and every core engineering course is substantially predicted by performance in some foundational courses, above and beyond the predictive role of general student academic performance mea-

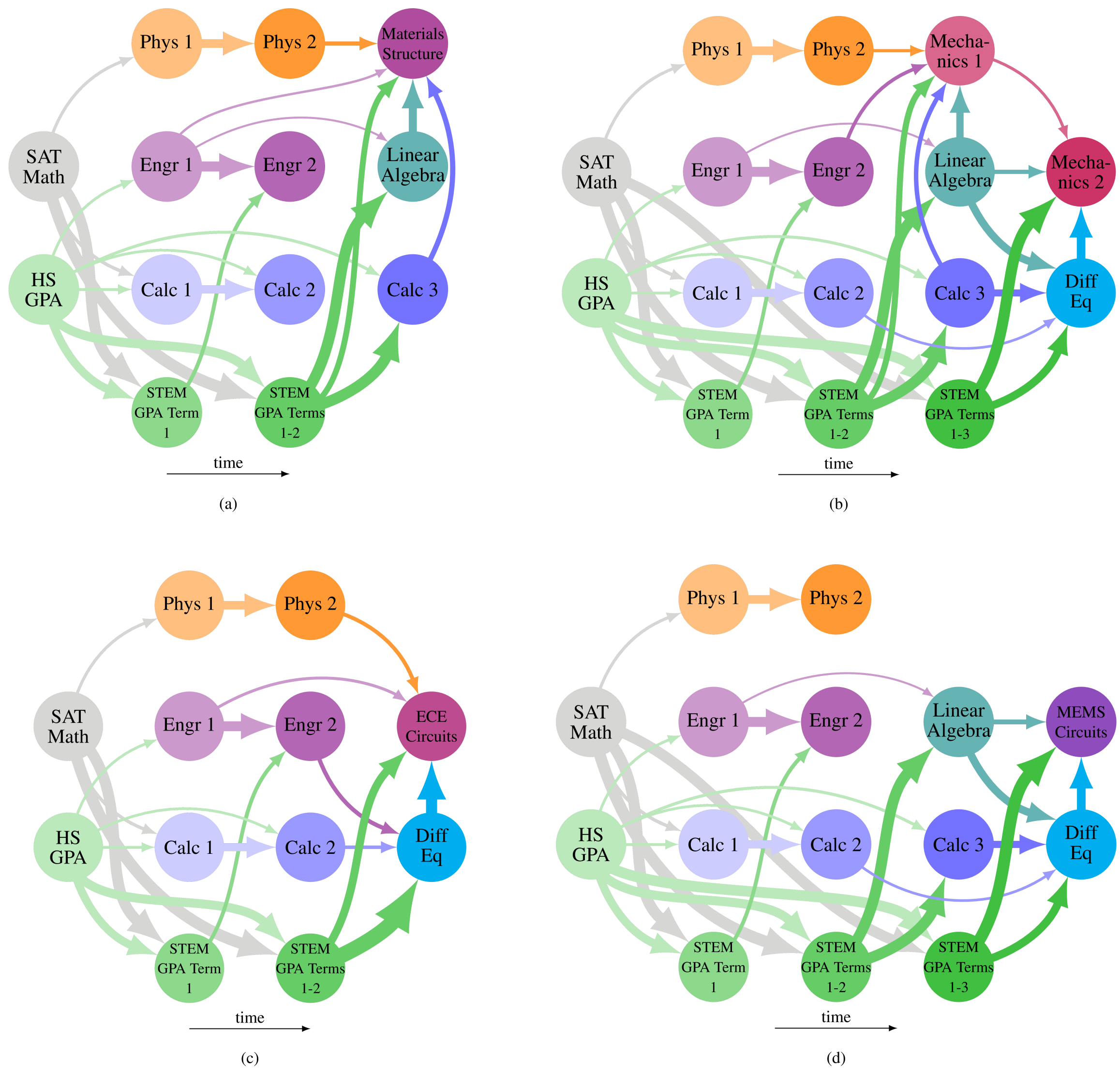


Figure 7: A visual representation of all statistically significant predictive relationships leading to various second year engineering courses: (a) Materials Structure and Properties, (b) Mechanics 1 and 2, (c) ECE Circuits, and (d) MEMS Circuits. Courses are organized left to right according to the chronology of the MEMS and EE curricula. Line thicknesses are scaled directly by β. Note that since prior STEM GPA values are used directly in the calculation of future STEM GPA values, the prior values are not included as predictors of later values.

sures. There is an important predictive role for such a general factor, which is best captured by the STEM GPA and to a smaller extent HS GPA. Indeed, chemistry courses gave spurious connections to courses that have no obvious chemistry content in them until STEM GPA was included in the analysis, while the predictive power of introductory physics and engineering is only slightly reduced. But the cumulative predictive strengths of the foundational courses are large in every case, above and beyond the general factor: sometimes the top foundational course was a stronger predictor than STEM GPA and always the combination of the top two foundational courses were a stronger predictor than STEM GPA (these trends are most easily seen in Table 11 by comparing the regression coefficients).

Turning to **RQ3** about which courses were most important, we find that university mathematics appears to be a general foundational pillar for engineering students in all their coursework. This result is consistent with the findings of Sadler and Tai [235] that high school mathematics is the general foundational pillar for university students' performance in introductory biology, chemistry, and physics. Note also that different core engineering courses depended upon different mathematics courses. In other words, it was not the case that performance in any single mathematics course is a good general indicator of student skill; rather different core engineering courses depend upon specific kinds of mathematics (Calculus 3, Linear Algebra, or Differential Equations).

The contrast of ECE vs. MEMS Circuits courses (lower two diagrams of Fig. 7) provides an interesting case because the content of these courses and the implementations are very similar according to their syllabi, but the sequence of prerequisite courses in the two majors is different. MEMS Circuits is taken one semester later, after students have taken Linear Algebra. For students in MEMS Circuits, Linear Algebra is a substantial predictor in addition to Differential Equations, which is taken concurrently. This suggests that the students in ECE Circuits may be at a disadvantage for not having had Linear Algebra. In order to test this, we conducted an ANCOVA on ECE Circuits grades based on the order in which Circuits and Linear Algebra were taken (three groups, Linear Algebra before, concurrent, and after or never); although there is a typical order, some students take Linear Algebra earlier than do others. It is important to note that course order is inherently confounded by academic skill; that is, stronger students are more likely to take second-year mathematics courses earlier. Therefore, it is important to control for such prior differences in this analysis. One particularly relevant proxy for mathematical skill is SAT Math score; controlling for SAT Math and high school GPA, we still see a significant benefit in ECE Circuits grade for those who took Linear Algebra prior to or concurrent with ECE Circuits, $F(2, 960) = 4.09$, $p < 0.02$.

Furthermore, these engineering students also rely on their introductory engineering courses in which they learn to use computational tools of engineering beyond mathematics, such as MATLAB, C++, and Python. Again, it is not that one of these serves as a general indicator variable for all core courses. Sometimes Engineering 1 was the significant predictor (i.e., for Material Structures and Linear Algebra in the MEMS curriculum and ECE Circuits in the ECE curriculum) and sometimes Engineering 2 was the significant predictor (i.e., Mechanics 1 in the MEMS curriculum and Differential Equations in the ECE curriculum); which of the two courses remains significant may depend upon what tools the later courses use.

Finally, these engineering students rely on core courses that are directly related to their discipline. For the physics-oriented engineering disciplines MEMS and ECE, that additional foundation is their introductory physics sequence. For example, doing well in the ECE Circuits is unlikely when students obtain Cs in Physics (as highlighted in Fig. 5), even when controlling for mathematics ability and general student ability.

4.4 General Discussion

Consistent with the finding of Sadler and Tai [235] for the transition from high school to college science courses, we find that mathematics is a foundational pillar of success for engineering students as they progress through their engineering curriculum. Our findings are not a simple replication, since the population of our study (engineering students in their first two years of college) is different than the population in Sadler and Tai's study (a variety of college STEM students in their transition from high school to college) and the target courses are different. Further, the variation in which form of mathematics is most important across engineering courses and the existence of predictive power above and beyond general STEM GPA predictors provide even stronger evidence for the role of mathematics course content rather than just a general mathematical ability factor. Consider, for example, the results for the fourth-semester course Mechanics 2 in Table 11 and Fig. 7b, where we delineate the various correlations of the third-semester courses Mechanics 1, Linear Algebra, and Calculus 3 to Mechanics 2. Notably, Calculus 3 is not a significant predictor while Linear Algebra is, showing that the predictive powers measured are sensitive to more than just the discipline and recency of the predictor course. Additionally, Linear Algebra is a stronger predictor of performance in Mechanics 2 (indicated by a higher β coefficient) than the prior course in the sequence, Mechanics 1, despite the fact that another mathematics course, Differential Equations, co-occurs as an even stronger predictor of Mechanics 2 while no other engineering courses appear as direct predictors. The current findings provide evidence that the details of the mathematical skills and knowledge per se are the foundation of the transfer.

Beyond these relationships between mathematics and engineering, we further found that physics was also a pillar of success for these MEMS and ECE students, albeit with a smaller effect than the concurrent mathematics courses. This relationship between foundational courses in physics and mathematics and subsequent courses in engineering at the college level is fundamentally different from the relationships observed by Sadler and Tai in that they saw no interdisciplinary effect: other than the transfer effects for mathematics, only within-discipline effects were observed. The difference between their findings and ours may partly be due to the depth of knowledge acquired in high school vs. in college and partly to the interdisciplinary, applied-science nature of engineering. This work provides evidence for course specific knowledge and skills transfer effects from foundational courses to later engineering courses. Our findings suggest that these foundational courses offer invaluable contexts for further analysis of transfer. For example, future research could examine particular knowledge and problem-solving skills that students transfer from the foundational courses to subsequent engineering courses from one context to another and what aspects of the curriculum facilitate or hinder such transfer of knowledge.

4.4.1 Implications for Instruction and Future Research

Our analysis using large institutional data at a large US-based research university validates the inclusion of these foundational courses in engineering curricula by revealing the strength of transfer of knowledge and skills from foundational courses to core courses using a multiple linear regression analysis. Similar analyses should be conducted to examine other engineering departments. While mathematics and physics were statistically significant predictors of performance in engineering disciplines that have a strong emphasis on mathematics and physics, we hypothesize that a similar investigation into chemical engineering curricula is likely to reveal predictive power of introductory chemistry for the subsequent engineering performance of those students. Further, we hypothesize that investigations of this nature by engineering departments at colleges and universities outside of the US would be similarly able to analyze how well their courses build upon one another. This is true for departments with curricula similar to the one described here – with early interdisciplinary courses followed by discipline-specific courses – or even more broadly for curricula which specialize earlier (e.g., in UK-based departments).

This type of large-scale investigation using institutional data to examine the predictive relationships between courses can play a central role in advising engineering students. For example, advisors guiding students through their first two years in an engineering program could benefit from a research-backed understanding of the course relationships in the engineering curricula as well as course performance trends, such as those seen in Fig. 5. We urge advisors using results such as these in a positive manner: to encourage students who have enrolled in these foundational courses to take them seriously to promote improvement in course grades, instead of discouraging students from pursuing an engineering major. Also, making note of these strong course relationships can help advisors identify potential indicators for a need for additional support. However, these types of additional support should be provided with careful planning in a manner that does not profile students, e.g., by potentially offering the same additional support to all students even though some students are particularly encouraged to take advantage of them.

Research-based evidence from the type of analysis presented in this investigation could be used to counteract problematic common wisdom being circulated among students about some foundational courses being unimportant to later success and that they should only strive to get passing grades. Similarly, departments seeking to (re)-design their curricula can also greatly benefit from this type of investigation because they need to understand the affordances and constraints involving relationships between different courses in the engineering curriculum that most strongly drive student performance. It is logical to wonder what the consequences of replacing these early courses with alternate course designs (e.g., by adding more intense design-project courses) would be; however, our findings suggest that there is a benefit to each of the current courses and it would therefore be unwise to replace the current courses without testing whether the replacement courses provided additional benefits.

Another equally important finding is the nature of STEM GPA as a good predictor of performance in later semesters. The β coefficients in Table 11 show that STEM GPA is a stronger predictor in semester 4 courses (MEMS Circuits, Mechanics 2) than semester 3 courses (ECE Circuits, Materials Structure, Mechanics 1). It is important to note that the follow up analysis on the VIFs indicated that STEM GPA is somewhat collinear with the other predictors, though this is expected given how it is calculated. That being said, the time when the STEM GPA is most closely related to the other predictors is in predicting term 2 courses, and STEM GPA is either a non-significant or, in the case of Engineering 2, a weak predictor of performance in term 2 courses. It is only in terms 3 and 4 that the STEM GPA variable becomes a significant and large predictor, perhaps in part because over time it becomes an increasingly sensitive measure relative to individual course grades. Regardless, this trend indicates that in later semesters, students are increasingly likely to perform similar to the way they have performed in the past. In particular, student mobility from a low GPA to high GPA is limited and students who typically obtain C grades in earlier courses are predicted to continue to perform at a similar level in later courses.

This lack of mobility brings up an important issue of equity since students who are less privileged are more likely to have had less than ideal K-12 education and are more likely to have lower grades in college [276]. Ideally, a college academic setting should provide appropriate guidance, mentoring, and support to close the academic opportunity gap and promote growth to ensure that all students develop similar levels of high competency regardless of their performance in the first year. Students who had strong and weak high school GPAs then initially obtained As and Cs, respectively, in the first year. If these two groups continue to perform on parallel trajectories throughout their college education and their performances in the later years do not become comparable, it may signal that the higher education institution may not be doing enough to close this opportunity gap over the years the students are in college engineering programs and ensure that all engineering students have the opportunity to excel and thrive, not just survive. Otherwise, students who complete their engineering degree with a low GPA are unlikely to find good career opportunities that are comparable to those offered to students with high GPAs.

In addition to wider efforts to make education equitable, advisors can play a critical role in supporting students who may be struggling. We again emphasize the importance of avoiding profiling students, which can be accomplished by giving the same advice to all students instead of only to a particular subset. Such advice could draw attention to the importance of students' foundational coursework to success in later engineering coursework; that is, if there are strong ties between courses and students get behind on the prerequisites, it can hinder learning throughout their undergraduate STEM major. This is particularly relevant in light of prior research showing that engineering students perceive first-year courses as unimportant to their overall academic career [96, 289, 67, 281].

Finally, we note that although the relationships reported here by analyzing large institutional data are inherently correlational, we have intentionally designed the selection of data for our regression analyses to enforce the chronology of the MEMS and ECE curricula in order to strengthen the validity of these predictive relationships and we have controlled for high school GPA and SAT Math scores, the strongest likely confounds in this kind of research. However, the inclusion of factors beyond course performance, e.g., survey data measuring engineering students' motivational beliefs about physics, mathematics, and engineering, could further strengthen these claims by providing additional context for what students believe about their own abilities in these various foundational subjects. Moreover, the addition of motivational characteristics as predictors could help isolate how academic performance can be partially based on motivational beliefs rather than entirely based on cognitive skills and content knowledge [178, 268]. Such analyses are critically important for evaluating how equitable current instructional systems are for underrepresented students who may be at increased risk, e.g., for anxiety due to stereotype threats. This type of information can further help one to contemplate strategies for designing equitable and inclusive learning environments. In particular, if correlations are found between these motivational characteristics and academic performance or pathways through a curriculum, then efforts to promote equitable and inclusive learning environments can also include strategies for boosting all students' motivational beliefs because there is a strong tie between these beliefs and student learning [285].

4.5 Conclusions

In this study, we investigated the relationships between grades earned in foundational courses and second-year engineering courses for students in two large engineering majors. We found that course grades in foundational courses relate linearly to grades in later engineering courses, indicating that simple linear regression models were appropriate for such prediction analyses. These regression models revealed that student performance in specific prior courses can be connected to performance in second-year engineering and mathematics courses, even when controlling for several general student performance factors, namely high school GPA, SAT Math score, and cumulative STEM GPA. Further, we found that advanced mathematics courses (i.e., Calculus 3, Linear Algebra, and Differential Equations) along with cumulative STEM GPA, were the strongest predictors of student performance in second-year engineering courses.

4.6 Acknowledgments

This research is supported by NSF Grant DUE-1524575 and the James S. McDonnell Foundation Grant 220020483. We are grateful to Larry Shuman for helpful discussions and feedback on this research.

5.0 Engineering Curriculum - Structural Equation Modeling

5.1 Introduction and Theoretical Framework

Engineering schools are increasingly recognizing the importance of evidence-based approaches to improve student learning to ensure that all students have the opportunity to excel regardless of their background [39, 37, 38, 138, 247, 255]. Holistic consideration of how these engineering programs are currently succeeding in supporting their undergraduate majors is crucial in order to make appropriate changes to the curricula and pedagogies based upon metrics informed by data and ensure that all students are adequately supported. Data analytics can provide valuable information that can be useful in making informed decisions and transforming learning for all students [6, 217].

Information obtained from data analytics on large institutional data in these areas can be an important component of understanding the role that foundational courses, e.g., in math and science, play in later engineering performance and determining whether there are gender differences in course performance or course relationships. Such results can aid in contemplating strategies for improving student support and ensuring that learning environments are equitable and inclusive so that all students can thrive. In order to improve diversity, it is important that engineering schools take a careful look at the extent to which their programs for the majors are equitable and inclusive and provide adequate support to all students, especially for groups that are historically underrepresented in STEM, namely women and students from diverse ethnic and racial backgrounds [199].

The theoretical framework for this research is based upon many prior studies investigating the relationships between course grades and various academic outcomes such as choice of major, retention in engineering programs, and degree attainment [21, 92, 94, 112, 189, 202, 203]. For example, one multi-institution collaboration leveraging institutional data for this type of research is the Multiple-Institution Database for Investigating Engineering Longitudinal Development (MIDFIELD) [165, 167, 211]. At the studied university, all engineering programs require students to complete a common first-year curriculum before beginning coursework for their chosen engineering major. Our study reported here uses 10 years of institutional data at a large research university to focus on these types of relationships during a critical transition for engineering majors, i.e., the transition from the first to the second year when these students choose their engineering major and begin taking courses specific for their chosen major. To our knowledge, no prior investigation in engineering education has focused on these specific issues which are critical for enhancing engineering education.

Our theoretical framework is further inspired by much prior research showing how gender differences in academic performance and career decisions of engineering students can result from biases and stereotypes [44, 45, 166, 170, 192, 210, 212, 222]. Many interrelated factors influence women's decision to pursue an education in engineering as well as subsequent decisions about which engineering major to study and even whether to remain in engineering [62, 63, 244, 125, 95, 21]. These factors include sociocultural factors, motivational factors, and various aspects of prior education such as quality of teaching [244, 63, 22, 243, 124, 260, 274, 4, 70, 94]. In particular, it has been proposed that cultural bias and stereotypes can negatively impact the self-efficacy and academic performance of women in various STEM subjects including mathematics and physics [63, 4, 70, 94]. This is potentially damaging to prospective women in engineering since success in mathematics and science courses in high school has a positive impact on students' choice of and persistence in an engineering major [274, 161] and longer-term career goals [292, 24]. Further, success in first-year college STEM courses continues to provide feedback to engineering students and inform their motivational characteristics [177, 178].

In order to gain an understanding of these issues central to improving education for all engineering majors, this research harnesses data analytics applied to 10 years of institutional data for Mechanical Engineering and Materials Science (MEMS) majors at a large state-related university to investigate how well the performance of MEMS majors in early foundational courses predicts performance in subsequent engineering courses. The MEMS curriculum was chosen for this investigation because the foundational courses in mathematics, physics, and chemistry are likely to be very important for these students to excel in their studies. The MEMS curriculum is also ideal to explore since mechanical engineering is the engineering program with the largest number of students at the studied university. We note that materials science undergraduate majors, though very few in number at the studied university, are also included in this analysis since they take the same courses considered in this study as mechanical engineering majors through the first two years. These courses for the majors have been offered for decades under the assumption that the later courses would build on the earlier ones coherently to help the majors build a robust knowledge structure and develop their problem solving, reasoning, and meta-cognitive skills. Investigating the predictive relationships between these courses will not only allow us to measure how well the courses in the curriculum build upon one another, but will also provide a structure in which we can test for gender differences throughout the model, both in grades earned and the strength of the relationships between courses.

This investigation can be useful for other institutions who may perform similar analyses in order to contemplate strategies for improving education in different engineering programs in a holistic manner as well as improving equity and inclusion. In particular, institutions could compare their findings with the 10-year baseline data provided here from a large state-related university for the synergy observed between courses in a mechanical engineering curriculum, gender differences throughout the curriculum, or use this analysis as a template for similar analyses of other engineering curricula.

5.1.1 Research Questions

Our research questions regarding the curriculum for MEMS majors at a large state-related university are as follows.

RQ1. Are there gender differences in course performance among MEMS majors in first-year foundational courses and second-year courses for the major?

RQ2. Does performance in first-year foundational courses and advanced mathematics courses predict performance in second-year MEMS courses?

RQ3. Does the degree to which earlier course grades predict later course grades differ for men and women?

5.2 Methodology

5.2.1 Measures

Using the Carnegie classification system, the university at which this study was conducted is a public, high-research doctoral university, with balanced arts and sciences and professional schools, and a large, primarily residential undergraduate population that is full-time and reasonably selective with low transfer-in from other institutions [133]. De-identified data were provided by the university on all engineering students who had enrolled in introductory courses from Fall 2009 through Spring 2019. The data include demographic information such as gender, which is central to this study. We note that gender is not a binary construct. However, the university data includes "gender" as a binary categorical variable. Therefore, that is how the data regarding gender are represented in these analyses. From the full sample of undergraduate engineering majors, a sub-sample was obtained by applying several selection criteria to select out MEMS majors from other engineering majors who took some MEMS courses listed in Table 12 (e.g., bioengineering, chemical engineering, and industrial engineering students all also take Mechanics 1). In particular, in order to be kept in the sample, students were required to meet the following criteria: 1) enroll in at least one of the two introductory engineering courses listed in Table 12, 2) enroll in Mechanics 2. Note that all of the courses we consider in this analysis in Table 12 are required courses in the curriculum for MEMS majors. After applying the selection criteria, the sample contains 1485 students. The students in the sample are 16.4% female and had the following race/ethnicities: 82.9% White, 7.2% Asian, 3.8% African American, 2.2% Latinx, and 3.8% Other or Unspecified.

The data also include high school GPA on a weighted 0-5 scale that includes adjustments to the standard 0-4 scale for Advanced Placement and International Baccalaureate courses. Finally, the data include the grade points and letter grades earned by students in each course taken at the university. Grade points are on a 0-4 scale with A = 4, B = 3, C = 2, D = 1, F = 0, where the suffixes '+' and '−' respectively add or subtract 0.25 grade points (e.g. B− = 2.75), with the exception of A+ which is reported as the maximum 4 grade points.

Table 12: All required first-year courses along with second-year mathematics and selected MEMS courses are listed. Full course names are given along with shortened names used elsewhere in this paper and the terms in which the courses are typically taken by MEMS majors.

Term	Full Course Name	Short Name
1	General Chemistry for Engineers 1	Chem 1
	Intro to Engineering Analysis	Engr 1
	Basic Physics for Science and Engineering 1	Phys 1
	Analytic Geometry and Calculus 1	Calc 1
2	General Chemistry for Engineers 2	Chem 2
	Intro to Engineering Computing	Engr 2
	Basic Physics for Science and Engineering 2	Phys 2
	Analytic Geometry and Calculus 2	Calc 2
3	Analytic Geometry and Calculus 3	Calc 3
	Intro to Matrices and Linear Algebra	Lin Alg
	Materials Structure and Properties	Mat Structure
	Statics and Mechanics of Materials 1	Mechanics 1
4	Differential Equations	Diff Eq
	Statics and Mechanics of Materials 2	Mechanics 2

5.2.2 Analysis

In order to evaluate the grades that MEMS majors earn in their courses by gender, we grouped students by the gender variable and computed standard descriptive statistics (mean, standard deviation, sample size) separately for each group [98]. Gender differences in course grades were evaluated using Cohen's d to measure the effect size [201, 195], as is common in education research [206].

The extent to which performance (i.e., grades earned) in earlier foundational courses predicts performance in later MEMS courses was evaluated using Structural Equation Modeling (SEM) [152]. SEM is the union of two statistical modeling techniques, namely Confirmatory Factor Analysis (CFA) and Path Analysis [152]. The CFA portion tests a model in which observed variables (or "indicators") are grouped into latent variables (or "factors"), constructed variables that represent the variance shared among all indicators that load on a particular factor [152]. The strength of the relationship between indicators and factors is measured by the factor loading, λ. Further, one of the unstandardized factor loadings per factor is fixed to $\lambda = 1$ in order to define the units and scale of the factor itself [152]. The factor loadings of other indicators are measured relative to this fixed factor loading [152]. The degree to which each indicator is explained by the factor is measured by standardizing the factor loadings, to $0 \leq \lambda \leq 1$, where λ^2 gives the percentage of variance in the indicator explained by the factor [152].

The Path Analysis portion then tests for the statistical significance and strength of regression paths between these factors, simultaneously estimating all regression coefficients, β, throughout the model [152]. This is an improvement over a multiple linear regression model in which only a single response (target or outcome) variable can be predicted at a time, which problematically disallows hierarchical structures [267]. By estimating all regression paths simultaneously, all estimates are able to be standardized simultaneously, allowing for direct comparison between standardized β coefficients throughout the model. In these models, we further estimate the intercepts (i.e., the mean when controlling for all predictors) of all indicators and factors. Indicator intercepts are denoted by τ and factor intercepts by α.

In this paper, we report the model fit for SEM using the Comparitive Fit Index (CFI), Tucker-Lewis Index (TLI), and Root Mean Square Error of Approximation (RMSEA) [152, 128]. Commonly cited standards for goodness of fit using these indices are as follows: For CFI and TLI, Hu and Bentler [128] found that many authors [128, 51, 229] suggest values above 0.90 and 0.95 indicate a good fit and a great fit, respectively. For RMSEA, several authors [128, 48] suggest that values below 0.10, 0.08, and 0.05 indicate a mediocre, good, and great fit, respectively.

Finally, these model estimations can be performed separately for different groups of students (e.g., men and women) using multi-group SEM. These differences are measured in a series of tests corresponding to different levels of "measurement invariance" in the model, with each step fixing different elements of the model to equality across the groups and comparing to the previous step via a Likelihood Ratio Test (LRT) [152]. A non-significant p-value at each step indicates that the estimates are not statistically significantly different across groups. "Weak" measurement invariance is demonstrated by fixing the factor loadings to equality, "strong" invariance is demonstrated by further fixing to equality the indicator intercepts, and finally "strict" invariance is demonstrated by further fixing to equality the residual error variance of the indicators. If measurement invariance holds at least through "strong" invariance, then all remaining differences between the groups occur at the factor level, either as differences in factor intercepts or β coefficients [152]. If instead measurement invariance does not hold, then the equality constraint on estimates (especially factor loadings and indicator intercepts) between groups can be relaxed for one estimate at a time in order to find the set of estimates for which partial measurement invariance holds. That is, the equality constraint can still be imposed on a subset of the factor loadings, intercepts, and/or residual variances, with the remaining estimates allowed to differ between groups.

Using SEM, we model student progression through the second year of the MEMS curriculum by grouping courses together into factors by their subject (e.g., introductory physics or advanced mathematics). Further, we found that all courses taken in the first year covary to such a degree that an overall first-year factor that loads on each of the first-year subject factors produces the best model fit. We use multi-group SEM to test for gender moderation, i.e., to test for gender differences in the predictive relationships in the model, as well as mean differences in course grades (indicators) and course factors [152].

Due to the nature of institutional grade data, modeling students' progress through an entire curriculum involves a large amount of missing data due to various reasons. These can include students receiving credit for courses taken elsewhere (e.g., over the summer at a different college), not completing the curriculum, skipping courses that are normally required with special permission, and the inevitable errors that occur in large datasets. The default approach to missing data in many modeling programs, listwise deletion, is then not desirable since it leaves very few students in the sample and can bias the results [204]. Considering this, we employed Full Information Maximum Likelihood (FIML) in order to impute missing data within the SEM model [152].

In addition to the aforementioned benefits of using SEM such as simultaneous estimation of all model elements and the ability to use FIML for missing data estimation, the basic structure of SEM also provides benefits to the modeling process. In particular, by first using CFA to group indicators into factors and then performing path analysis on those factors, the effect of measurement error is minimized since the error variance will be left at the indicator level and does not contribute to the estimation of regression coefficients at the factor level [152].

All analyses were conducted using R [226], making use of the package `lavaan` [234] for the SEM analysis and the package `tidyverse` [279] for data manipulation and descriptive statistics.

5.3 Results

5.3.1 Gender Differences in Course Performance

In order to investigate for gender differences in course grades and answer **RQ1**, we grouped students by the gender variable and first calculated the standardized mean difference, Cohen's d, to measure the effect size of the gender differences [201, 195]. Table 13 shows these results for all MEMS students who at least continued through Mechanics 2 (typically taken in the fourth term). Note that since enrollment in Mechanics 2 was used as a selection criterion, the population in every other course is less than that in Mechanics 2, since students may be missing grades for previous courses for a variety of reasons. We find that, on average, women performed similar to or slightly better than men in all courses except introductory physics (Physics 1 and 2). This general pattern matches that of high school GPA, though the effect sizes of the gender differences in the courses is small, with the highest difference occurring in Linear Algebra a small effect size ($d = 0.24$), and a medium effect size in high school GPA ($d = 0.47$). Though still small in effect size, the gender differences seen in introductory physics are the only ones in which men earn higher grades on average, despite the same population showing women performing better than men in high school GPA and grades in other courses.

Looking at the patterns of gender differences in Table 13 by subject shows that in the second year (terms 3 and 4), this pattern of women earning higher grades than men is still present in advanced mathematics and is on par with the strongest gender differences observed in the first year. On the other hand, in their MEMS courses, men and women are earning more similar grades, except women earning slightly higher grades in Mechanics 1.

The full grade distributions as described on average in Table 13 are shown in Fig. 8 (first-year courses) and Fig. 9 (second-year courses). In these distributions, we can see that in most courses, especially the second year courses in Fig. 9, women earn A and A+ grades at a slightly higher rate than men, who in turn have a slightly higher rate in earning lower grades. As noted in the preceding discussion of Table 13, physics was the only subject with appreciable grade differences favoring men, and that can be seen again in Fig. 8, with men earning higher rates of grades B+ and higher, and women earning higher rates of B grades and lower.

In some courses where we saw very little mean gender differences in Table 13, we can still observe some interesting grade distributions. For example, Calculus 2 in Fig. 8 has an alternating pattern of men and women earning higher rates of the various letter grades. Further, the course with the gender difference closest to that of high school GPA, namely Linear Algebra in Table 9, has a noticeably large rate of A grades earned by both men and women, but especially large for women. Finally, the courses in both Figs. 8 and 9 display a general trend of peaks of varying sizes at A, B, and C grades with the exception of Engineering 1, which shows an especially high mean with a single peak at A.

5.3.2 Predictive Relationships Between Courses

Turning then to **RQ2**, we use SEM to test for the degree to which performance in earlier courses predicts that of later courses in the curriculum. The full 1485 student sample was used in all stages of SEM, with FIML employed to impute missing data [152]. We grouped first-year courses by their subject (Calculus, Chemistry, Engineering, and Physics), then further grouped these four subjects into a "First Year" factor. Second-year mathematics courses are grouped together into an "Advanced Math" factor, and MEMS courses are left ungrouped in order to separately predict the grades in each of these courses, as well as to allow the term 3 courses (Mechanics 1 and Materials Structure) to predict the term 4 course (Mechanics 2).

The final model is shown in Fig. 10 (CFI = 0.972, TLI = 0.968, RMSEA = 0.042, all indicating a great model fit [128, 51, 229, 48]), in which non-significant regression paths have been trimmed from the model. Note that since there are gender differences present, all values shown are the unstandardized values, for which a majority of the factor loadings, intercepts, and regression coefficients have been fixed to equality [152]. The primary flow of predictive paths is such that each first year subject loads strongly on the overall First Year factor. This First Year factor then strongly predicts the Advanced Math factor, which in turn strongly predicts each of the MEMS courses. There are two smaller additional regression paths, with Chemistry predicting Materials Structure and Materials Structure predicting Mechanics 2, both over and above the primary predictive paths from Advanced Math.

5.3.3 Gender Differences in the Structural Equation Model

To test for gender differences and answer **RQ3**, we used multi-group SEM to test for differences in the model [152], first testing factor loadings, then indicator intercepts, then residual variances, and finally regression paths. The estimates that differed for men and women are reported in Fig. 10. In each step, the model fit was great, with CFI $>$ 0.95, TLI > 0.95, and RMSEA < 0.05 [128, 51, 229, 48]. We did not find full measurement invariance at either the factor loading ("weak") or item intercept ("strong") stages [152]. That is, when fixing all factor loadings to equality, the Likelihood Ratio Test (LRT) showed significant differences in the model with $p < 0.05$ unless some estimates were allowed to vary between men and women [152]. In particular, Fig. 10 shows that the factor loading (λ) of Physics on the First Year factor is slightly lower for women ($\lambda_{\mathrm{F}} = 0.84$) than for men ($\lambda_{\mathrm{M}} = 0.89$). Though seemingly a small difference, this gender difference in factor loading is sufficient to account for men's higher performance in physics courses overall (Table 13 and Fig. 8). This relationship will be explored more carefully in the next section. Further, three courses showed differences in their intercepts (τ): Physics 1 ($\tau_{\mathrm{F}} = 0.62$, $\tau_{\mathrm{M}} = 0.71$), Engineering 2 ($\tau_{\mathrm{F}} = 1.02$, $\tau_{\mathrm{M}} = 1.16$), and Linear Algebra ($\tau_{\mathrm{F}} = 0.63$, $\tau_{\mathrm{M}} = 0.51$). Each of these indicates deviations from the course grade gender differences that would be predicted solely by the difference in high school GPA and, in the case of Physics 1, the aforementioned factor loading difference.

Notably, Fig. 10 shows that apart from the mean difference in high school GPA ($\tau_{\mathrm{F}} = 4.07$, $\tau_{\mathrm{M}} = 3.88$), there are no additional gender differences present in any regression paths leading to calculus, chemistry, or MEMS courses. This does not mean that the model predicts no gender differences in these courses. Rather, this means that the gender differences observed in these courses are consistent with the gender difference observed in high school GPA propagating through the model's predictive paths to each course. That is, the women MEMS majors are coming in with a higher high school GPA than men and earning higher grades than men in the majority of their courses consistent with that high school GPA difference. The exceptions to this occur only in four courses, with the largest departure from high school GPA occuring in the physics sequence, and especially in Physics 1 (in which the students learn mechanics).

5.3.4 Model-Implied Gender Differences in Individual Courses

For the courses that are grouped into factors, we can calculate the model-implied gender differences from high school GPA using the underlying equations governing the structural equation model. The grades earned in each of those courses, G_{c}, which are only directly predicted by their corresponding subject factor, η_{s}, are given by

$$G_{\text{c}} = \tau_{\text{c}} + \lambda_{\text{c}}\eta_{\text{s}} + \varepsilon_{\text{c}}, \tag{1}$$

where τ_{c} is the course grade intercept, λ_{c} is the course factor loading on the subject factor, ε_{c} is the residual variance in G, and η_{s} is the value of the subject factor. In this and subsequent equations, the variance terms are used to account for individual differences in grades earned and will ultimately be set to 0 in order to consider the predicted average grades. Each subject factor is in turn predicted by the overall First Year factor, η_{FY}, either through a factor loading or a regression path, and the First Year factor is in turn predicted by high school GPA, G_{HS}. So each subject factor can be written as

$$\eta_{\text{s}} = \alpha_{\text{s}} + \lambda_{\text{s}}\eta_{\text{FY}} + \zeta_{\text{s}} \tag{2}$$

or, in the case of advanced math which has a regression path rather than a factor loading from the First Year factor,

$$\eta_{\text{math}} = \alpha_{\text{math}} + b_{\text{math,FY}}\eta_{\text{FY}} + \zeta_{\text{math}}, \tag{3}$$

where in both cases ζ is the residual variance of η_{s} and α_{s} is the factor intercept. Similary, η_{FY} can be expressed in terms of G_{HS} which, since it has no predictors, is simply its mean plus its variance:

$$\eta_{\text{FY}} = \alpha_{\text{FY}} + b_{\text{FY,HS}}G_{\text{HS}} + \zeta_{\text{FY}} \tag{4}$$

$$G_{\text{HS}} = \tau_{\text{HS}} + \varepsilon_{\text{HS}}. \tag{5}$$

In order to consider the average model-implied gender differences, we will set all $\varepsilon = 0$ and $\zeta = 0$, since those variances account for the variability in individual students' grades. Further, the α of each factor is set to 0 by default in SEM. Considering all of this, combining all of the equations and setting the appropriate values to 0 yields

$$G_{\text{c}} = \tau_{\text{c}} + \lambda_{\text{c}}\lambda_{\text{s}}b_{\text{FY,HS}}\tau_{\text{HS}}, \tag{6}$$

where in the case of advanced math λ_{s} is replaced by $b_{\text{math,FY}}$.

Gender differences in the model shown in Fig. 10 occur in three places in this equation, namely τ_{HS} for all courses, λ_{s} for physics courses, and τ_{c} for Physics 1, Engineering 2, and Linear Algebra. Defining the gender difference in any given variable x as

$$\Delta x = x_{\mathrm{F}} - x_{\mathrm{M}} \tag{7}$$

then finally allows us to write the model-implied gender difference in these courses as

$$\Delta G_{\mathrm{c}} = \Delta\tau_{\mathrm{c}} + \lambda_{\mathrm{c}} b_{\mathrm{FY,HS}} \Delta\left(\lambda_{\mathrm{s}}\tau_{\mathrm{HS}}\right), \tag{8}$$

where we note that $b_{\mathrm{FY,HS}}$ does not vary by gender, nor does λ_c for any course.

Table 14 shows these values calculated for all first-year courses and advanced mathematics courses. A similar exercise can be carried out for the MEMS courses, making sure to account for multiple predictive pathways for Materials Structure and Mechanics 2. Note that these may differ from the observed gender differences in Table 13 due to statistical error introduced by the equality constraints placed on the estimates (i.e., constraining a value to be equal for men and women) as well as fixing all variances to 0. Nonetheless, an explicit calculation of these model-implied gender differences in individual courses is useful in examining which areas of the model account for the gender differences.

In particular, Table 14 shows that the τ_{c} difference observed in Engineering 2 in Table 13 is nearly exactly countering the predicted gender difference favoring women via τ_{HS}. In Linear Algebra, the τ_{c} gender difference indicates a course-specific boost for women similar in magnitude to that from τ_{HS}. And finally in Physics 1 and Physics 2, even though women have a higher mean high school GPA, the observed λ_{s} difference is sufficient to entirely eliminate the higher mean high school GPA for women and predict higher average grades for men in physics overall, with an additional τ_c difference in Physics 1 specifically.

5.4 Discussion

In response to **RQ2**, the only research question that did not deal with gender differences, the SEM diagrammed in Fig. 10 shows an overall pattern of strong cohesion through the first two years of this MEMS curriculum. In particular, the strong predictive pathway from high school GPA to the First Year factor to the Advanced Math factor to each MEMS course shows a robust cohesion throughout this curriculum. Further, we see that the First Year factor itself is loaded on very strongly by each of the first-year courses, though notably least strongly by first-year engineering courses (which consists of a sequence focusing on introduction to programming for engineers and other engineering basics), perhaps due to the high grades overall and narrow grade distributions in those courses.

Turning then to **RQ1**, by looking at the average grades earned by men and women in these courses in Table 13, we find that women tended to earn similar or slightly higher grades on average than men in most subjects. This pattern is consistent with their average high school GPA, where these same women ($\tau_{\mathrm{F}} = 4.07$) have a slightly higher GPA than the men ($\tau_{\mathrm{M}} = 3.88$). The only subject that does not fit this pattern is physics, in which, on average, the men are earning higher grades despite having a lower high school GPA and women earning higher grades in every other concurrent course. We do find that the gender differences in the courses have, at most, a small effect size ($|d| \leq 0.24$ for all courses in Table 13), despite a medium effect size in the high school GPA gender difference ($d = 0.47$). However, we find that in physics, instead of the gender gap merely diminishing relative to high school GPA similar to other college courses, women are now earning lower grades than men, inconsistent with high school GPA and all other course grades (where women always earn higher or comparable grades compared to men). This could be indicative of strong stereotype threats due to societal biases associated with physics as well as an inequitable or non-inclusive environment in these introductory physics courses that is disproportionately negatively affecting women [46, 86, 292, 251].

Testing these gender differences further using multi-group SEM in order to answer **RQ3** provided further support for these interpretations of Table 13. In nearly all courses, the gender differences in high school GPA were sufficient to explain course grade gender differences. In two individual courses there were small additional gender differences found, namely Engineering 2 in which women and men earned the same grades on average (and so the course-level gender difference slightly favors men controlling for high school GPA), and Linear Algebra in which women earn slightly higher grades than otherwise predicted. These single-course differences may simply be normal noise or may have an underlying cause, but neither fits a larger pattern. On the other hand, introductory physics shows a gender difference affecting both Physics 1 and Physics 2 (namely, the λ difference from the First Year factor to the Physics factor) in addition to an intercept difference in Physics 1. In all of the gender analysis for **RQ1** and **RQ3**, the introductory physics sequence is standing out as a source of gender differences in this curriculum which may be due to inequity.

We note that this model focuses on the relationships between the grades earned and does not explicitly account for different ways in which gender disparities in early coursework can affect students (e.g., through self-efficacy, sense of belonging, etc.). Grades earned play a key role in students' crucial decisions about whether to remain in college and which major to pursue [24, 274, 161, 292]. In particular, one mechanism by which this occurs is the feedback loop between course grades and self-efficacy [292, 8, 9, 10, 224, 214, 46]. Other studies at this same university have found significant gender differences favoring men in the physics self-efficacy of students in large introductory physics courses [177, 178], consistent with studies at other universities [53, 239]. These findings suggest the need for efforts towards improving equity and inclusion, particularly in introductory physics courses, including interventions designed to boost students' self-efficacy, growth mindset and sense of belonging in physics [280, 270, 271, 33]. Due to these relationships between grades, self-efficacy, and enrollment decisions, the first year of any curriculum is an especially vital time to ensure that all students have a high sense of belonging and self-efficacy.

We also note that our analysis focused on those students who, when entering the second-year, chose to major in a "physics-heavy" engineering major and still we observed these gender differences surrounding introductory physics. A future study could investigate whether the gender differences in course grades in introductory physics are even more pronounced among the students in less physics heavy engineering majors such as bioengineering, chemical engineering, or industrial engineering.

Finally, SEM can be used to evaluate the cohesiveness of a curriculum by examining the predictive relationships between course grades for students with different demographics backgrounds, including gender, race and ethnicity, low socioeconomic status, and first-generation college students. Further, course grade differences can be evaluated at the same time across the full curriculum, and how later gender differences are explained by earlier gender differences in the curriculum can provide insight into the best practices to address such disparities and improve equity and inclusion in the learning environment.

5.5 Acknowledgments

This research is supported by the National Science Foundation Grant DUE-1524575 and the Sloan Foundation Grant G-2018-11183.

Table 13: Descriptive statistics are reported for grades in courses taken by MEMS majors through the second year, on a 0-4 scale, and high school GPA on a weighted 0-5 scale. Only students who have taken Mechanics 2 are reported in order to restrict to MEMS majors. Reported are the sample size (N), mean grade points earned (μ), and standard deviation of grade points (σ) for men and women separately, along with Cohen's d measuring the effect size [201, 195] of the gender difference. $d < 0$ indicates the mean for men is higher, $d > 0$ indicates the mean for women is higher.

Course	Gender	N	μ	σ	d
High School GPA	F	243	4.07	0.37	0.47
	M	1242	3.88	0.44	
Chemistry 1	F	191	2.76	0.80	0.09
	M	933	2.68	0.90	
Chemistry 2	F	185	2.67	0.82	0.19
	M	873	2.51	0.83	
Engineering 1	F	187	3.63	0.40	0.18
	M	937	3.55	0.45	
Engineering 2	F	197	3.28	0.65	-0.01
	M	978	3.28	0.65	
Physics 1	F	192	2.79	0.68	-0.15
	M	927	2.90	0.72	
Physics 2	F	201	2.67	0.73	-0.08
	M	1040	2.73	0.80	
Calculus 1	F	140	3.16	0.68	0.18
	M	719	3.03	0.70	
Calculus 2	F	170	2.90	0.84	0.01
	M	871	2.89	0.84	
Calculus 3	F	228	2.94	0.86	0.14
	M	1157	2.81	0.92	
Linear Algebra	F	222	3.26	0.79	0.24
	M	1184	3.06	0.88	
Diff Eq	F	235	3.01	0.83	0.19
	M	1204	2.84	0.95	
Mat Structure	F	220	2.92	0.89	0.01
	M	1204	2.92	0.93	
Mechanics 1	F	243	3.25	0.74	0.08
	M	1236	3.19	0.78	
Mechanics 2	F	243	2.94	0.95	0.00
	M	1242	2.94	1.03	

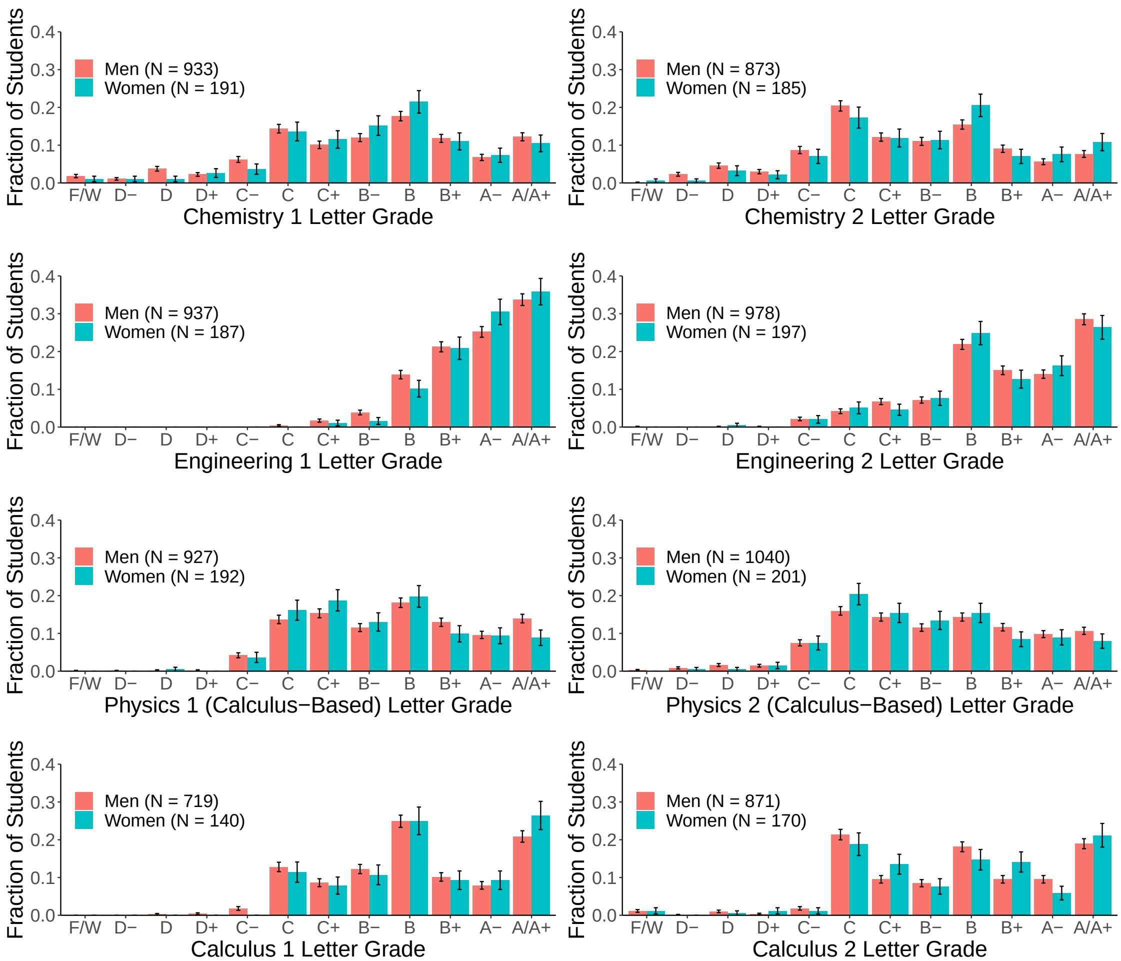


Figure 8: Grade distributions of MEMS majors in first-year courses, plotted separately for men and women. The proportion of each gender group that earns each letter grade is plotted along with the standard error of a proportion [98].

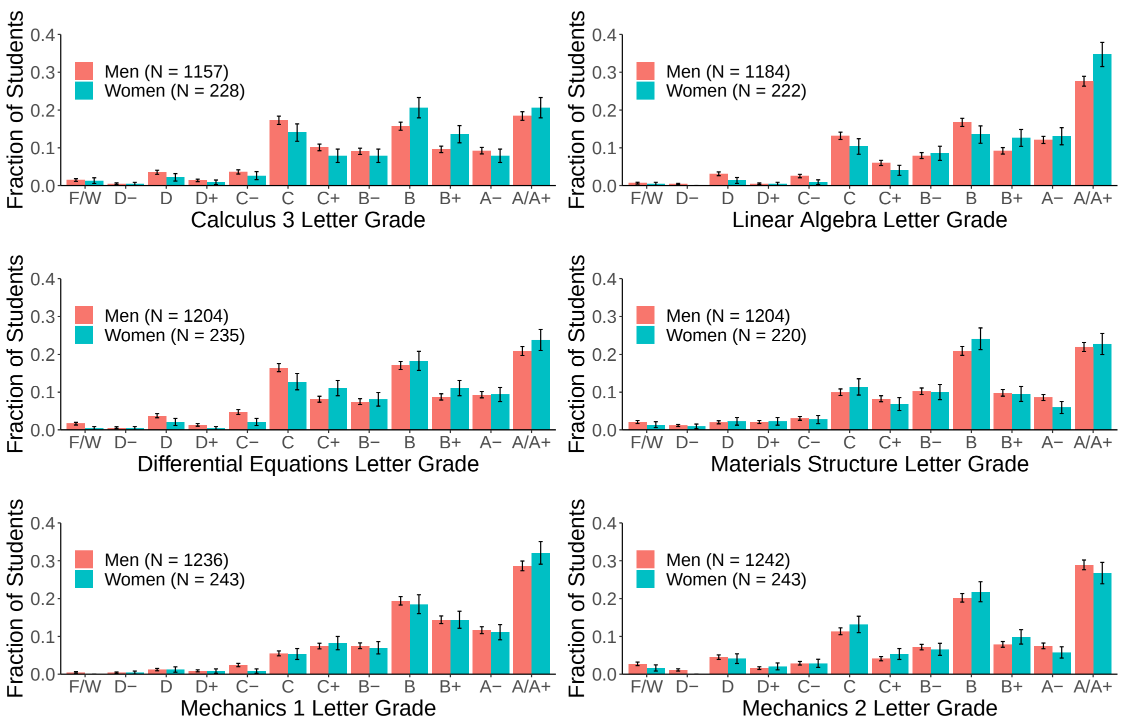


Figure 9: Grade distributions of MEMS majors in second-year mathematics and engineering courses, plotted separately for men and women. The proportion of each gender group that earns each letter grade is plotted along with the standard error of a proportion [98].

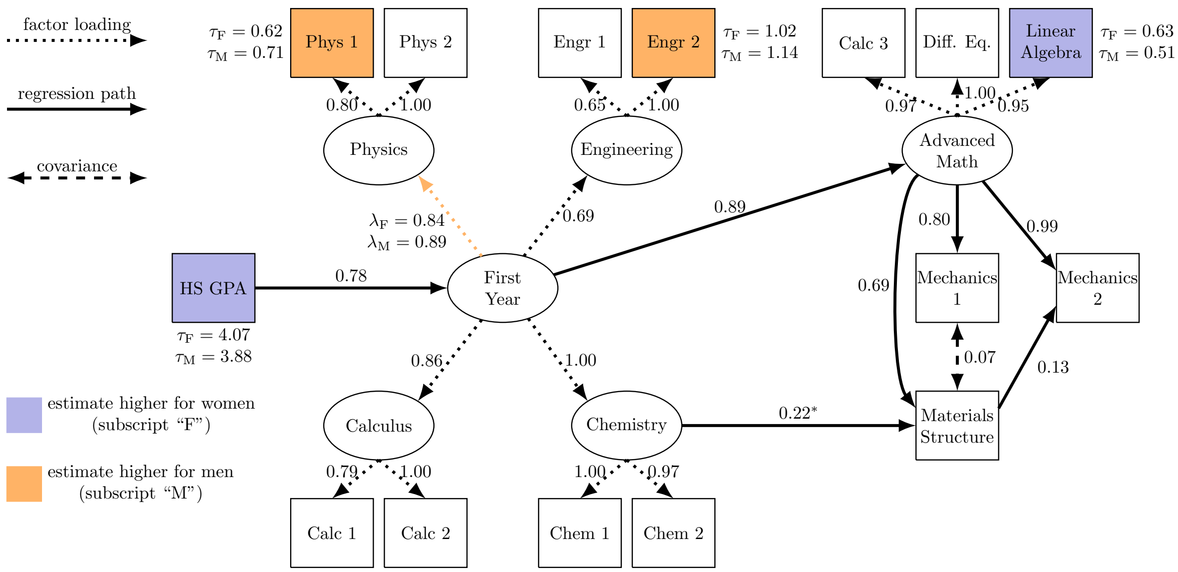


Figure 10: A diagram of the SEM model designed to test for the relationships between courses in the MEMS curriculum, as well as gender differences therein. All 1485 students in the sample were included in the model, with FIML used to estimate missing data. Reported next to each line are the unstandardized values for factor loadings, regression coefficients, and covariances. Estimates that differ between female (subscript "F") and male (subscript "M") students are reported separately for each gender, in this model these are differences in intercepts (τ) and factor loadings (λ). High School GPA (HS GPA) and all first-year factors were allowed to regress on Advanced Math and the second-year engineering courses, but many paths were not statistically significant ($p < 0.05$) and thus are not shown. All drawn paths are significant to the $p < 0.001$ level except the one denoted with a superscript $*$, which is significant to the $p < 0.01$ level. All missing paths are not statistically significant, with $p > 0.05$. Line styles indicate the type of relationship between connected items, with factor loadings, regression paths, and covariances represented by dotted, solid, and dashed lines, respectively (please see legends on the left hand side for details).

Table 14: Model-implied average gender differences in first-year foundational courses and advanced mathematics courses, each column separately rounded to two decimal places. Reported are the gender differences in each term of Eq. 8 as well as the full course grade gender difference (ΔG_{c}, which is the sum of the previous two columns). In particular, the $\Delta\tau_{\mathrm{c}}$ column represents gender differences in individual courses that are not accounted for via the subject factors, and the $\lambda_{\mathrm{c}} b_{\mathrm{FY,HS}} \Delta\left(\lambda_{\mathrm{s}}\tau_{\mathrm{HS}}\right)$ column represents gender differences accounted for via the initial high school GPA gender difference propagating through the First Year and subject factors in Fig. 10.

Course	$\Delta\tau_{\mathrm{c}}$	$\lambda_{\mathrm{c}} b_{\mathrm{FY,HS}} \Delta\left(\lambda_{\mathrm{s}}\tau_{\mathrm{HS}}\right)$	ΔG_{c}
Chemistry 1	0.00	0.15	0.15
Chemistry 2	0.00	0.15	0.15
Engineering 1	0.00	0.07	0.07
Engineering 2	-0.11	0.10	-0.01
Physics 1	-0.09	-0.04	-0.13
Physics 2	0.00	-0.05	-0.05
Calculus 1	0.00	0.10	0.10
Calculus 2	0.00	0.13	0.13
Calculus 3	0.00	0.13	0.13
Linear Algebra	0.12	0.13	0.24
Diff Eq	0.00	0.13	0.13

6.0 Comparison of Self-Efficacy and Performance of Engineering Undergraduate Women and Men

6.1 Introduction

Self-efficacy is intertwined with many aspects of academic life, so understanding both the nature of self-efficacy itself and its effects is crucial in order to improve education. Self-efficacy is the belief in one's capability to succeed in a particular task or subject [9, 10, 214], and in an academic context it can both affect and be affected by academic performance, a feedback loop that can either promote or hinder student learning [8, 292, 224, 46]. Most importantly here, the feedback loop can produce growing inequities for those traditionally underrepresented in engineering [228, 191]. Due to this recursive nature, measuring self-efficacy and academic performance longitudinally is vital to understanding their relationship. We present such a longitudinal analysis of the relationships between self-efficacy and academic performance using institutional data and survey responses from one US-based institution to test for patterns of gender differences. While the study is implemented within one US-based engineering program, the broader methodology described can be applied at any institution in order to seek sources of gender inequities inherent in their academic programs.

6.1.1 Underrepresentation of Women in Engineering Fields

The overall underrepresentation of women in Science, Technology, Engineering, and Mathematics (STEM) careers is largely due to their large underrepresentation in engineering [199]. Progress has been made in some fields of engineering, but the largest engineering majors remain heavily male-dominated [199]. Many interrelated factors influence women's decision to pursue an education in engineering as well as subsequent decisions about which subfield to study and even whether to remain in engineering [95, 125, 62, 63, 244]. These factors include sociocultural factors, motivational factors, and various aspects of prior education such as quality of teaching [63, 244, 4, 70, 94, 260, 124, 22, 273, 243]. In particular, it has been proposed that cultural bias and stereotypes can negatively impact the self-efficacy and academic performance of women in various STEM subjects such as mathematics and physics [244, 4, 70, 94]. This is potentially damaging to prospective women in engineering since success in mathematics and science courses in high school has a positive impact on students' choice of and persistence in an engineering major [45, 170, 222, 227, 161, 274] and longer-term career goals [24, 292].

6.1.2 Discipline-Specific Variation of Gender Differences in Self-Efficacy

Much of the research on self-efficacy has examined broad discipline self-efficacy (e.g., science or engineering self-efficacy) or more discipline-specific self-efficacy (e.g., physics or chemistry self-efficacy), especially as students become more specialized in tertiary education (e.g., become chemistry majors vs. physics majors). Surprisingly, however, there has been little research comparing self-efficacy in various disciplines among the same population.

This gap in research comparing discipline-specific self-efficacy is particularly problematic for engineering. First, large differences in self-efficacy by gender have been previously found in various STEM disciplines in which engineering majors typically take courses [125, 62, 22, 42, 129, 53, 239, 177, 268]. These self-efficacy differences have been implicated in gender differences in performance, retention, and choice of major [161, 274, 24, 210, 292], which is especially relevant during the students' first year. There are also hints that the size and even direction of self-efficacy differences vary by academic discipline. For example, women have equal or slightly higher self-efficacy in English [216] and equal self-efficacy in biology [215, 266]. Second, it is difficult to generalize gender patterns in general-education STEM courses taken by many non-engineering majors to engineering students in particular because there are large variations in which students choose to take the various general STEM courses and because engineering students are potentially different from the general science major population (e.g., women choosing to major in the male-dominated engineering may have atypically high self-efficacy beliefs in physics and math relative to the general science major population). Third, variation in gender differences in self-efficacy by discipline in later years might be explained by differential rates of participation in engineering majors by gender (e.g., male-dominated electrical and mechanical engineering vs. more balanced chemical engineering may reflect differential gender differences in physics vs. chemistry self-efficacy).

Previous work has shown that women tend to have a lower self-efficacy than men in disciplines including physics [53, 239, 177] and chemistry [268] as well as mathematics and engineering [125, 62, 22, 42, 129]. However, these studies of physics, chemistry, and mathematics self-efficacy did not focus on engineering students.

6.1.3 Time-Varying Gender Differences in Self-Efficacy

Another important gap in the self-efficacy research of particular relevance to engineering education is the lack of research on the change in student self-efficacy across the full course of studies in engineering. The self-efficacy differences that were in place at the end of the degree will be most relevant to those later career transitions. As course grades are generally higher in the more advanced courses [153] and experience with success accumulates, gender differences initially observed in the earlier years may disappear. On the other hand, due to the negative effects of having low self-efficacy on exam performance, the feedback loop from performance to self-efficacy, and the potentially negative effect of being a numerical minority in most courses (at least within some engineering majors), early gender differences could actually magnify over time.

A few studies have investigated self-efficacy changes over the course of two semesters, and generally found little change in self-efficacy over that shorter time period [157, 177, 178]. One study found that over two years, the self-efficacy of women in engineering showed a positive trend [176]. Little is known about gender differences in self-efficacy of graduating engineering students.

6.1.4 Gender Differences in Alignment of Self-Efficacy and Academic Performance

There are multiple factors that lead to alignment of self-efficacy beliefs and academic performance. First, a central source of self-efficacy beliefs is prior performance feedback [8, 9, 10, 11, 12, 13]. Second, self-efficacy can influence academic performance [8, 9, 10, 11, 12, 13, 46, 224, 292] Third, factors that directly influence self-efficacy can also directly influence academic performance. In particular, stereotype threats that women experience in many STEM disciplines due to societal stereotypes and biases can increase their anxiety, rob them of cognitive resources while solving problems, and lead to reduced test scores [46]. Additionally, these stereotype threats can lower self-efficacy, which can result in reduced interest and engagement during learning [86, 292].

However, these factors may influence performance and self-efficacy in different manners, so academic success may not be fully aligned with self-efficacy. Of particular relevance here, alignment between academic performance and self-efficacy may be differential by gender. Women in STEM may interpret struggle in difficult courses as being due to inability whereas men may interpret struggle in these same courses as due to lack of effort [251]. Further, gendered stereotypes, differential availability of role models, and consistently being in a numerical minority in coursework may produce biased self-efficacy.

6.1.5 Research Questions

The current study examined gender differences in longitudinal measurements of student self-efficacy and academic performance in math, science, and engineering courses. Our work extends previous research by investigating trends within the same populations (engineering overall and by separate engineering majors) across four different STEM disciplines in order to situate any gender differences observed within the broader context of these students' academics. This research is of particular importance to engineering given the poor gender diversity in engineering overall and the negative role that low self-efficacy can play in students' academic decisions [95, 125, 62, 63, 244].

Our research questions to investigate the trends in self-efficacy and academic performance are as follows:

RQ1. Do men and women's self-efficacy in various core disciplines change along different trajectories as they progress from their first to their fourth year?

RQ2. Do gender differences in performance within foundational STEM courses vary by course discipline?

RQ3. Is there a match between gender differences in self-efficacy and gender differences in performance?

RQ4. Does being in a male-dominated engineering major change the gendered trajectories in self-efficacy relative to students in an more gender-balanced major?

6.2 Methodology

6.2.1 Participants

Using the Carnegie Classification of Institutions of Higher Education [133], the US-based university at which this study was conducted is a public, high-research doctoral university, with balanced arts and sciences and professional schools, and a large, primarily residential undergraduate population that is full-time, more selective, and lower transfer-in. De-identified demographic data and university course grade data were provided by the university on all first-year engineering students who had enrolled from Fall 2009 through Spring 2018. We recognize that gender is not a binary construct; however, the data provided lists gender only as a binary categorical variable, so we present our analyses and results accordingly. Since all of our analyses will involve gender, we have filtered out students from the sample whose gender is unknown since they would later be omitted for each analysis.

The sample of engineering students for whom we have gender and grade data consists of 3,928 students. A subset of this sample also participated in surveys administered by the School of Engineering from Spring 2014 through Spring 2017 at the end of their first, second, and/or fourth years. The average response rate for each year was 79%. The full sample of students was 27% female and had the following race/ethnicities: 80% White, 8% Asian, 5% African American, 2% Latinx, and 5% Other. The mean age at the beginning of the students' first year was 18.9 years (SD=1.7 years), reflecting a population of students who predominantly are attending college immediately after completing high school.

6.2.2 Measures

6.2.2.1 Grades The data provided include the grade points (GPs) earned in all courses at the university, the semester and class in which the course was taken, and the grade point distributions (mean and standard deviation) for each class. GPs are on a 0-4 scale (F= 0, D= 1, C= 2, B= 3, and A= 4) where '+' and '–' suffixes add or subtract 0.25 (e.g., B+= 3.25) except for A+, which is recorded with a GP of 4.

6.2.2.2 Declared Major We also have the declared major(s) for each student for every semester. At this university, engineering students initially choose an engineering major only at the end of their first year. Therefore, the first declared major is likely an accurate measure for most students. Possible majors are: Mechanical Engineering and Materials Science, Electrical and Computer Engineering, Chemical and Petroleum Engineering, Civil and Environmental Engineering, Bioengineering, or Industrial Engineering.

6.2.2.3 Discipline-Specific Self-Efficacy The self-efficacy data were collected as part of an online survey that the engineering school gives to all engineering students at the end of the spring semester of their first, second, and fourth years. Students are given a few reminders to complete the survey and are told that this survey is important for evaluating the effectiveness of the engineering program, resulting in a completion rate exceeding 75% and sometimes higher than 90%. The four analyzed items asked students to "Please rate your level of confidence in the following knowledge and skill areas: My ability to use my knowledge of [mathematics/engineering/physics/chemistry] to solve relevant engineering problems." It is considered best practice in the design of self-efficacy ratings to identify a particular task context to allow respondents to make reliable judgments [8]. The use of the phrase "to solve relevant engineering problems" in each survey question serves this purpose, in addition to increasing the relevance of the judgments to engineering education.

The students were given five options – "poor," "fair," "good," "very good," and "excellent" – recoded as 1 to 5. Although some survey scales produce non-interval data that should not be analyzed as interval data, the Likert rating scales for measuring self-efficacy typically produce normally distributed data, as they did in the current study (see Appendix B.1). Further, these self-efficacy scales are always analyzed as interval data (e.g., by computing means, using t-tests and linear regressions).

In smaller scale research studies, a survey scale often has multiple items so that scale reliability can be calculated (e.g., Cronbach's alpha). In larger scale longitudinal studies, survey fatigue becomes a major concern as students stop responding with having to repeatedly do long surveys, and it is not uncommon to use only one survey item per scale [49, 23]. In our case, the self-efficacy items were embedded in a larger survey and asking students three to five questions on self-efficacy per domain would have been received negatively. The main disadvantage of one-item scales is the increase in measurement noise (i.e., a problem of reliability, not a problem of validity), which reduces the ability to detect effects. This deficit is overcome by using a large sample, as in the current study.

To validate the use of these single-item scales, a subset of students ($N = 446$) also completed a multi-item physics self-efficacy survey within their Physics 2 class around the time that they were also completing the general engineering attitudinal survey, and the two measures were highly correlated ($r = 0.60$). Finally, the general physics self-efficacy survey showed a similar gender effect size (measured by Cohen's d) as the engineering-context physics self-efficacy ($d = 0.76$ in the physics context vs. $d = 0.84$ in the engineering context).

6.2.3 Analysis

Grouping by gender and major cluster. For the entirety of this analysis, students are grouped by their reported gender in order to investigate gender differences in the perception of and performance in foundational subjects in engineering. To investigate whether these gender differences differ across different engineering departments, we used the first major declared by these engineering students, which typically occurs at the end of their first year when the students move from the standard first-year courses to a specific engineering department's curriculum.

Students were grouped into three clusters of engineering majors determined by the proportion of women in those majors. Typically, these engineering students first declare their major at the end of their first year. The specific majors that went into each cluster were determined by the proportions of women and men who declared that major. Table 15 shows the engineering majors in each major cluster as well as the number of students and percentage of that number that are women, both for the full sample and the sample of survey takers.

6.2.3.1 Self-Efficacy Differences by Gender In order to test for statistically significant differences in self-efficacy scores, we performed *t*-tests comparing self-efficacy scores of men and women in engineering. Effect sizes of the gender differences were calculated in standard deviation units via Cohen's *d* [65]. Analyses of gender differences in self-efficacy were run separately for each of the four disciplines (mathematics, engineering, physics, and chemistry) at each time point (but averaging across the five cohorts of students). Furthermore, these tests were run for all available students first, then separately for the men and women in each major cluster.

6.2.3.2 Course Performance Differences by Gender Similarly, using *t*-tests [195, 201], we investigated gender differences in course performance on the grade points earned by men and women in each of the foundational courses. Again, the magnitudes of differences were calculated in standard deviation units. The courses we investigated were the foundational courses taken by the largest number of students in the School of Engineering, namely all of the common first-year courses in engineering, physics, chemistry, and mathematics as well as a selection of second-year mathematics courses taken by students in a variety of engineering departments: two in chemistry, two in physics, two in engineering, and five in mathematics. We note that the introductory engineering sequence at the studied university is a two-course sequence designed to teach the students computer programming skills in an engineering context, and in particular teaches the students to use MATLAB, C++, and Python to solve engineering problems. Further, we recognize that the content of such a course, or of the introductory curriculum as a whole, may vary from country to country and even institution to institution within a country. Thus, the investigation presented here

Table 15: The sample size of students in each major cluster, along with the percentage of women in that sample. Two samples are reported for each cluster, one for all students for which we have grade data and another for the subset for which we have survey data. Survey takers counted here may have taken any combination of the first year, second year, and fourth year surveys.

		All Students		Survey Takers	
Cluster	Disciplines	N	% Women	N	% Women
1	Electrical Computer Mechanical	2128	16%	991	22%
2	Chemical Environmental Civil	1551	34%	738	37%
3	Bioengineering Industrial	1185	41%	700	46%

provides a methodology for investigating these relationships within any particular curriculum, and the results presented here may be used as a comparison to a particular institution with a strictly enforced first-year engineering curriculum that includes courses in chemistry, engineering, mathematics, and physics.

6.3 Results and Discussion

6.3.1 Longitudinal Gender Differences in Self-Efficacy

In order to understand the perceptions of these engineering students about their foundational course work and answer **RQ1**, and specifically address how these perceptions differ for men and women, we plot in Fig. 11 the mean self-efficacy scores of men and women in engineering in each of the four foundational subjects (mathematics, engineering, physics, and chemistry) at each time point (end of the first, second, and fourth years). The full distributions of responses to these prompts are available in Appendix B.1. Looking at the data for the first year, there is a statistically significant gender gap favoring men in self-efficacy scores for applying mathematics, engineering, and physics to their work in engineering, and no difference in chemistry. Both mathematics and engineering follow a similar trajectory in that the initial gap remains in the second year and is eliminated by the fourth year. In sharp contrast, the gap shrinks but remains relatively large in physics even by the end of the fourth year. At no point is there a significant gender difference in chemistry self-efficacy.

Although we do not focus on this issue here, we note that self-efficacy of both men and women appears to grow, as expected, over years; the lack of growth in chemistry self-efficacy may reflect the relatively small role chemistry plays in the largest majors within the six engineering departments. Note that there is relatively little change in majors after students declare majors in their second years, nor is there much attrition overall in Engineering at this university after the second year. Thus, the changes between second and fourth years are unlikely to be caused by major switching or attrition. These arguments are further supported by follow-up analyses that included only data from students completed surveys at all three time points.

6.3.2 Performance Differences in Foundational Courses

Gender differences in performance were investigated across the foundational first year engineering curriculum and selected common second-year mathematics courses to answer **RQ2**. Table 16 reports the summary statistics (population N, mean μ, and standard deviation SD) for each course along with a p-value from a t-test comparing the grades earned by men and women in that course and the effect size (Cohen's d).

For all but one course, there were statistically significant gender differences, with all but one of those statistically significant results satisfying $p < 0.01$. Most interestingly, the direction of the differences varied by discipline. Only for the two introductory physics courses did men receive higher grades on average than women. In all the courses in the other three disciplines, women received higher grades on average than did men (and statistically significantly so except for Engineering 2). Moreover, although the lowest mean grades occurred in physics, the gender patterns in physics cannot be explained by physics being the most difficult course because students had similarly low grades in chemistry and calculus but with opposite gender differences (with women on average performing better than men in all chemistry and mathematics courses). Similarly, the differences could not be explained in terms of the stronger role of mathematics in physics versus chemistry because women had higher grades in every single mathematics course. Further, these gender differences in course grades could have implications for future success in engineering. Our previous study of engineering students' performance in these foundational courses at the same university found strong relationships between the grades earned in introductory courses and advanced mathematics courses, which in turn were statistically significant predictors of performance mechanical engineering and materials science courses [278].

It should be acknowledged, however, that none of the gender differences in course performance was large. Instead, what is surprising is the pattern of large gender differences in self-efficacy despite small differences in performance as well as performance and self-efficacy scores showing opposite trends (e.g., in mathematics and engineering, women on average

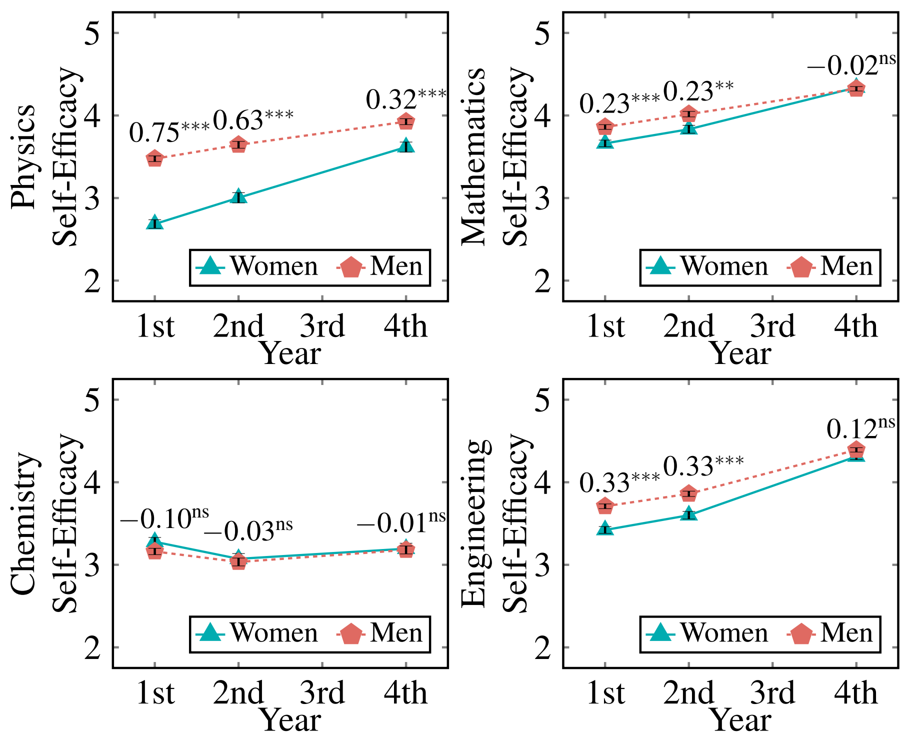


Figure 11: The mean self-efficacy scores of engineering students at the end of their first, second, and fourth years in each of the foundational subjects in engineering are plotted along with their standard error. Self-efficacy was measured on a Likert scale from 1 to 5. The vertical range of self-efficacy scores has been restricted to better show the gender differences. Above each pair of points, Cohen's d is reported (with $d < 0$ indicating women have a higher mean and $d > 0$ indicating men have a higher mean) along with the statistical significance of the gender difference according to a t-test, with $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$, and $^{\mathrm{ns}}p > 0.05$.

Table 16: Reported are the performance differences between female (F) and male (M) engineering students for grades earned in introductory courses in engineering, physics, chemistry, and mathematics as well as second-year mathematics courses. We report the sample size N, mean M, and standard deviation SD separated for men and women, as well as the p-value from a t-test comparing the grades earned and Cohen's d measuring the effect size, both for the individual courses and for each of the four subjects overall. The sign convention for Cohen's d matches that of Fig. 11.

Course	Gender	N	M	SD	p	Cohen's d Course	Cohen's d Subject
Engineering 1	F	1070	3.68	0.38	< 0.001	-0.19	-0.12
	M	2715	3.60	0.44			
Engineering 2	F	1095	3.33	0.66	0.168	-0.05	
	M	2797	3.30	0.72			
Physics 1	F	1127	2.60	0.82	< 0.001	0.19	0.14
	M	2824	2.75	0.86			
Physics 2	F	1123	2.59	0.85	0.015	0.08	
	M	2998	2.67	0.91			
Chemistry 1	F	1062	2.91	0.86	< 0.001	-0.13	-0.15
	M	2688	2.80	0.91			
Chemistry 2	F	1010	2.79	0.85	< 0.001	-0.17	
	M	2514	2.64	0.91			
Calculus 1	F	857	3.04	0.90	0.002	-0.12	0.15
	M	2199	2.92	0.96			
Calculus 2	F	989	2.87	1.00	< 0.001	-0.1	
	M	2563	2.74	1.08			
Calculus 3	F	1193	2.88	1.03	< 0.001	-0.12	
	M	2954	2.75	1.10			
Linear Algebra	F	762	3.30	0.91	< 0.001	-0.25	
	M	2296	3.05	1.06			
Differential Equations	F	1176	2.89	1.04	< 0.001	-0.14	
	M	3189	2.74	1.12			

have better grades but have lower self-efficacy than men). This contrast is directly examined in the next section.

6.3.3 The Relationship Between Self-Efficacy and Course Performance

In order to answer **RQ3** and investigate the relationship between self-efficacy and performance, we combined the two previous analyses for **RQ1** and **RQ2** to plot the effect sizes of gender differences in both self-efficacy and course grades (Fig. 12). For each point in the plot, we used only the subsample for which we had both a course grade (Ns varying from 579 for Linear Algebra to 1,163 for Engineering 1) and a self-efficacy score in the nearest survey (first year for the introductory courses, and second year for the second-year mathematics courses), a restriction which slightly alters the effect sizes from those of the full samples in Fig. 11 and Table 16. In addition to dashed lines along $d = 0$ on both axes, there is a dotted line along $d_{\mathrm{SE}} = d_{\mathrm{CG}}$ (where the effect size of self-efficacy differences equals the effect size of course grade differences), which represents where the data would fall if there was a one-to-one relationship between the effect sizes of self-efficacy and course grade. In addition, a vertical line is shown from the center of each discipline to the $d_{\mathrm{SE}} = d_{\mathrm{CG}}$ line, which represents the deviation of self-efficacy differences from academic performance differences.

Here we see the trends among the foundational subjects shine through strongly. As before, we see that engineering and mathematics are more similar than the other two disciplines. However, it is highly noteworthy that the directions of the self-efficacy and performance gender differences for all courses within two subjects are in direct opposition. In both engineering and mathematics, men have a higher self-efficacy, while women earn higher grades on average.

In Fig. 11, chemistry was the only subject in which there was no gender difference in self-efficacy, though we did see a performance difference favoring women. In Fig. 12, with the subset of the population for which we have both self-efficacy scores and course grade data, we see that chemistry falls right along the $d_{\mathrm{SE}} = d_{\mathrm{CG}}$ line. That is, for the populations completing the surveys, there is direct correspondence between self-efficacy and performance: women have higher self-efficacy in chemistry to the exact same extent to which they also

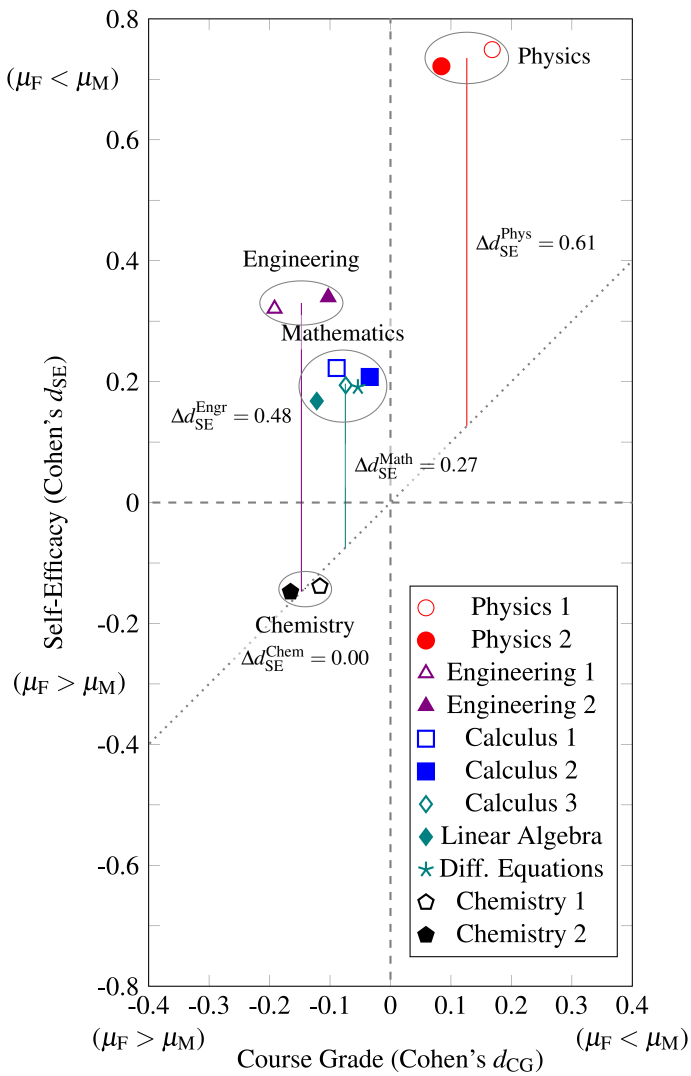


Figure 12: The effect sizes (Cohen's d, sign convention matching Fig. 11) of gender differences in self-efficacy (d_{SE}) and course grades (d_{CG}) are plotted for each of the introductory courses as well as second-year mathematics courses. Dashed lines for $d = 0$ on both axes as well as a dotted line along $d_{SE} = d_{CG}$ have been added. Ellipses group all courses in each subject. Each point contains the data of only those students for which both a grade and self-efficacy score were available. Introductory course grades (the first and second courses in a sequence) are paired with first-year self-efficacy scores, while second-year mathematics course grades are paired with second-year self-efficacy scores. Vertical lines have been added showing the distance along the self-efficacy axis from the center of each subject (defined as the average position of the constituent courses) to the $d_{SE} = d_{CG}$ line.

outperform men in chemistry.

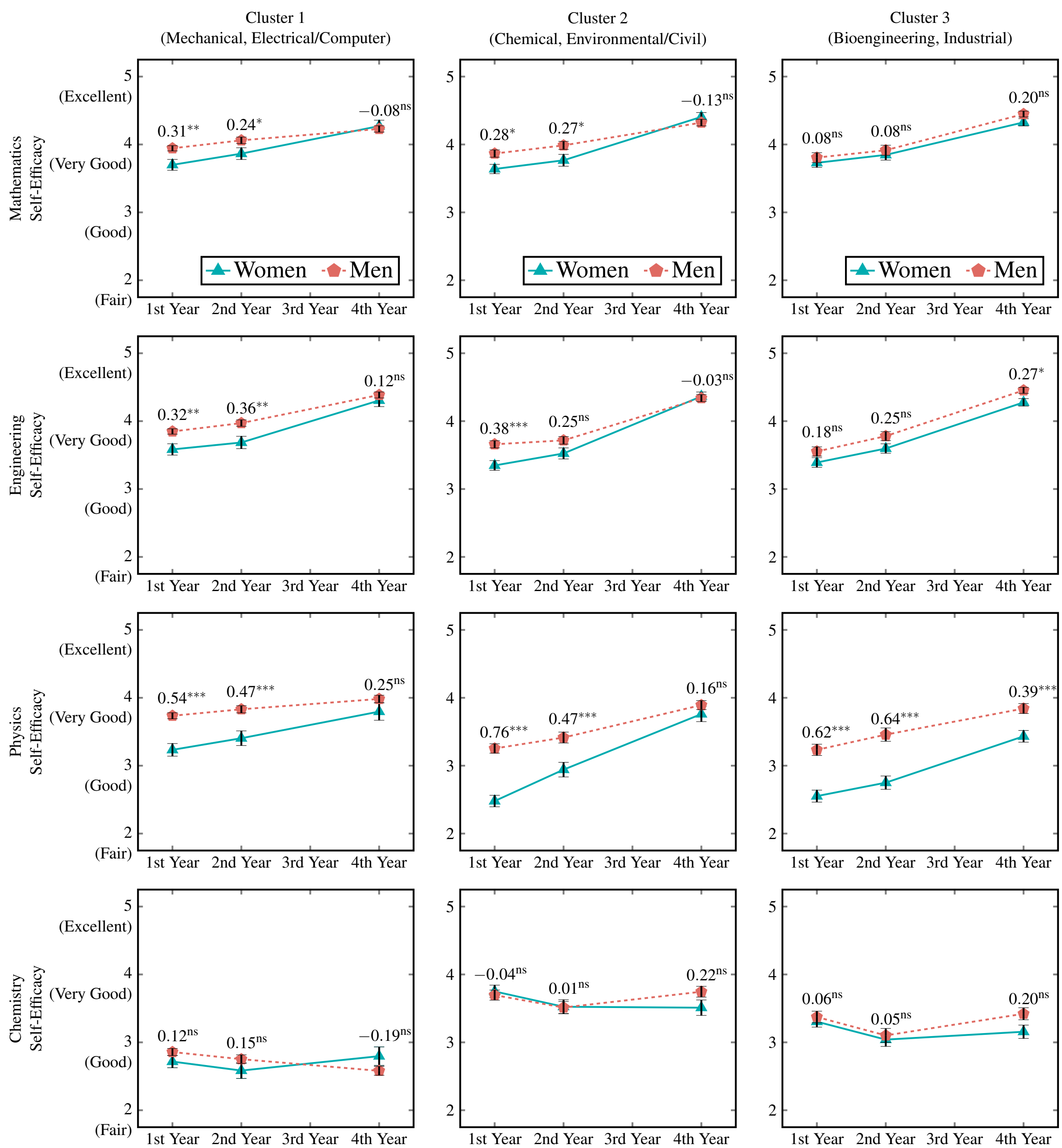
Figure 12 also shows an interesting contrast between quadrant and deviation from the matching self-efficacy/performance line. Both chemistry and physics are in 'match' quadrants of the figure, but while chemistry is precisely on the $d_{\mathrm{SE}} = d_{\mathrm{CG}}$ line, physics is not. On the one hand, men have higher self-efficacy in applying physics to engineering and do have higher performance, which is roughly a match. On the other hand, effect sizes are completely mismatched: the self-efficacy gap is much larger than the course grade gap leading to physics lying further away from the $d_{\mathrm{SE}} = d_{\mathrm{CG}}$ line than any other discipline.

6.3.4 Self-Efficacy Time Trends for Different Major Clusters

In Fig. 13, as in Fig. 11, we plot the mean self-efficacy scores of men and women in each of the four foundational subjects, but now separately for each major cluster. The three clusters of majors considered are 1) mechanical/materials science and electrical/computer engineering; 2) chemical and environmental/civil engineering; and 3) bioengineering and industrial engineering (see Table 15). The mean responses in each subject for each of the six majors separately are in Appendix B.2. Looking first at mathematics and engineering, we observe some small differences across the major groups, namely clusters 1 and 2 have self-efficacy differences in the first and second years that close by senior year while cluster 3 shows no self-efficacy differences in mathematics or engineering except a marginally significant gap in mathematics self-efficacy in the fourth year. As before there is a high degree of similarity between the responses in relation to mathematics and engineering within each cluster, suggesting that the strong tie in engineering students' perceptions of their abilities in mathematics and engineering is true for all engineering students, even despite overall differences between cluster 3 and clusters 1 and 2.

The third row of Fig. 13, showing responses for physics self-efficacy, shows a higher physics self-efficacy among students who choose physics-oriented majors such as electrical and mechanical engineering than students in other majors. Most importantly, Fig. 13 sheds additional light on likely causes of the unusually large and not closed gender gap in physics self-efficacy across majors. The cluster with the highest percentage of women (bioengineering/industrial engineering) is actually the only subset where the physics self-efficacy gap remains open by the fourth year. In other words, being in courses that are male dominant (at least by numbers) does not appear to be the central cause of the effect.

A related point has to do with ruling out simple exposure to physics content. If the



(Caption on next page.)

Figure 13: (Previous page.) As in Fig. 11, the mean self-efficacy scores of engineering students at the end of their first, second, and fourth years in each of the foundational subjects in engineering are plotted along with their standard error separately for students in each of the three clusters of majors. Self-efficacy was measured on a Likert scale from 1 to 5. The range of self-efficacy scores has been restricted to better show the gender differences. Each column contains the graphs for the different major groups while each row contains the graphs for self-efficacy in the different foundational subjects. Above each pair of points, effect size (Cohen's d, sign convention matching Fig. 11) is reported along with the statistical significance of the gender difference according to a t-test, with $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$, and $^{\text{ns}}p > 0.05$.

gender effect in physics were simply a result of a relative exposure to physics content, then we should expect that men in cluster 3 also have lower physics self-efficacy than students in cluster 1. While that is somewhat true for the first two years – possibly driven by a selection effect with the students that have the highest physics self-efficacy being most likely to choose a mechanical, electrical, or computer engineering major – that gap closes for men in cluster 3 by the fourth year, but not for women in the same majors. Further, it is noteworthy that the women in the physics-oriented majors in cluster 1 display a remarkably similar trend in physics self-efficacy to the men in clusters 2 and 3.

Finally, the fourth row of Fig. 13 shows the chemistry self-efficacy scores for each of the major groups, again showing a disparity based on chosen major. Cluster 2, which includes chemical engineering, has the highest chemistry self-efficacy across all years, followed by cluster 3, and finally cluster 1, students in which seem to display a decrease in their perceived to applying chemistry to engineering that persists through the fourth year. Despite having no gender gaps, Fig. 13 does show strong differences between the three major clusters, consistent with the idea that self-efficacy may play an important role in major selection (i.e., the highest self-efficacies occur from the start in students who selected chemical engineering). Turning to lower exposure/experience effects, Cluster 1's chemistry self-efficacy scores are the only ones which remain below 3 on the Likert scale even through their fourth year. Interestingly, a very small (albeit non-significant) gender gap in chemistry self-efficacy emerges in all groups by the fourth year, with men on average having a slightly higher chemistry self-efficacy than women in clusters 2 and 3 (including chemical engineering and bioengineering) and women on average having a higher chemistry self-efficacy than men in cluster 1. However, none of these differences by gender or time in chemistry self-efficacy were large.

6.4 General Discussion

RQ1 considers gender differences and temporal persistence of gender differences in engineering students' self-efficacy within four core disciplines. We observed no significant gender differences in chemistry self-efficacy in this undergraduate engineering population while that same population showed consistent differences in mathematics, engineering, and physics, with men tending to have a higher average self-efficacy in the first two years and the gap reducing by the fourth year. These effects were very small in mathematics and engineering, but large in physics. Given that a shared population of students was investigated (and they were all engineering students), this was the first clue that that are discipline-specific biases at play underlying gender differences in self-efficacy.

Turning to **RQ2**, we also observed consistent gender differences in course grades in each of the four core disciplines. Here however, women performed slightly better than men on average in mathematics, engineering, and chemistry. Such course differences were consistent with trends within this population for high school GPA differences by gender, and somewhat consistent with national trends [181], where mathematics shows mixed gender differences across institutions, although the same study finds men earn higher grades on average in chemistry. In contrast, men performed slightly better on average than did women in the two physics core courses that were examined. Thus, both self-efficacy and performance display substantial variation in gender differences by core STEM discipline, but not in consistent ways.

Focusing more specifically on alignment of self-efficacy with performance for **RQ3**, we observed three different patterns: 1) complete alignment with chemistry; 2) opposing direction small effects in mathematics and engineering; and 3) consistent direction but mismatching effect sizes in Physics (i.e., large self-efficacy effects but small performance differences). Thus, in answer to **RQ3**, gender differences in self-efficacy are inconsistently aligned with performance differences, strongly suggestive of some source of bias by gender in self-efficacy estimates that is discipline specific.

RQ4 investigated one possible source of such biases in self-efficacy: the experience of being a numerical minority in course work [228, 191]. Here the self-efficacy trends through the later years was most relevant because it is via those courses that the experience of engineering students varies by major. However, the data did not support an effect of being a numerical minority: the change patterns were generally similar across majors. The one exception actually went in the opposite direction: the large physics self-efficacy gender difference was primarily found in the majors with the highest proportion of women.

While these analyses are inherently correlational, the correlational pattern has ruled out some commonly offered explanations for gender differences in performance and self-efficacy. First, the self-efficacy gender differences cannot be a simple reflection of actual performance differences; if anything, the pattern is more consistent with self-efficacy biases causing performance differences. Second, performance differences in physics cannot be attributed to deficits in mathematical ability: among engineering students, women outperform men on average in every single mathematics course. Third, self-efficacy differences by gender do not seem to be driven by the experience of being a numerical minority in coursework.

6.4.1 Implications for Instruction and Future Research

The observed correlational findings have important implications for research and practice. We divide those implications around the sources of initial differences by discipline and then the differential change over the years by discipline. Focusing first on the initial differences, other work [53, 239, 177] has shown that engineers show large physics self-efficacy differences by gender even in the first few weeks of class in their first year. Physics tends to involve particularly strong beliefs that talent is required for doing well and the common cultural stereotype is that males are the ones that have such talent [77, 104, 173]. Regardless of the source of these beliefs, counter programming in secondary schools should be introduced to broadly address stereotypes and improve women's self-efficacy for physics (e.g., using growth mindset interventions [103, 285, 220]). If broadly distributed, the number of women applicants to engineering schools could rise. If more narrowly targeted to incoming engineering first-year students (a task more in the control of universities), prior biases might be lessened just in time to reduce stereotype-threat effects within the early coursework.

Turning to the differential changes in self-efficacy during the undergraduate years, at least the studied institution could perhaps be congratulated for having been able to reduce and sometimes entirely eliminate initial gender differences in self-efficacy. Future research should examine how broadly engineering programs around the world are able to achieve this outcome. Retention patterns at this university were not unusual by national US trends [199], which suggests that many US-based universities have also been similarly successful.

However, there are still important challenges: differences in self-efficacy remained overall, particularly in bioengineering and industrial engineering majors. Additional research should now focus on the experience of these students to understand why their physics self-efficacy remains so low. Physics is an important foundation to bioengineering overall and to many aspects of industrial engineering. Along these lines, replications at other universities are needed to examine whether these patterns are characteristics of those majors more generally or whether they come from department-specific messaging and coursework at this US-based university.

Another gap to acknowledge in the current study is the exclusive focus on gender and gender as a binary construct. Engineering is not only male-dominated: there is also underrepresentation by other gender identities and by sexual orientation, by race/ethnicity, and there can be additional effects at the intersection of race/ethnicity and gender (e.g., larger gender effects within underrepresented minorities) [262, 230, 5, 131]. The current study offers an analytic framework that could equally well be applied to better understand those other domains of underrepresentation, even if they potentially have different explanations.

6.4.2 Conclusions

Self-efficacy is a specific attitude that can play a strong role in influencing student performance in both the short-run (by undermining studying and exam performance) and in the long-run (by influencing degree persistence). In this study, we add important nuances to the common previous finding of lower self-efficacy by women in STEM. On a positive note, we observed reductions in the self-efficacy gap over time (**RQ1**), but the early gender gaps are still important because they can affect differential attrition, which is generally highest in the early years of the engineering degree. However, we find that the self-efficacy gender difference varies by specific STEM topic (**RQ2**), even within a particular sample of engineering students, with especially large differences in physics self-efficacy. Further, we also draw attention to the non-normativity of these self-efficacy gender differences: women sometimes have lower self-efficacy even when they have higher performance such as in engineering and mathematics courses (**RQ3**). Finally, we show that although there is some variation in these patterns by engineering major, there is no support for larger gender-based self-efficacy gaps being found in majors where women are a numerical minority (**RQ4**).

6.5 Acknowledgments

This research is supported by NSF Grant DUE-1524575, Alfred P. Sloan Foundation Grant No. G-2018-11183, and the James S. McDonnell Foundation.

7.0 An Intersectional Perspective on Enrollment Patterns

7.1 Introduction and Theoretical Framework

The importance of evidence-based approaches to improving student learning and ensuring that all students have the opportunity to excel regardless of their background is becoming increasingly recognized by Science, Technology, Engineering, and Mathematics (STEM) departments across the US [136, 137, 172, 39, 37, 38, 119, 72, 118]. With advances in digital technology in the past few decades, institutions have been keeping increasingly large digital databases of student records. We have now reached the point where there is sufficient data available for robust statistical analyses using data analytics that can provide valuable information useful for transforming learning for all students [6, 217]. This has lead to many recent studies utilizing many years of institutional data to perform analyses that were previously limited by statistical power [166, 167, 211, 181, 283]. Therefore, here we focus on harnessing institutional data to investigate the obstacles faced by students with various disadvantages who must overcome obstacles in their pursuit of higher education.

The theoretical framework for this study has two main foundations: critical theory and intersectionality. Critical theories of race, gender, etc. identify historical sources of inequity within society, that is, societal norms that perpetuate obstacles to the success of certain groups of disadvantaged people [69, 150, 286, 110, 258, 261, 242]. Critical theory tells us that the dominant group in a society perpetuates these norms, which are born out of their interests, and pushes back against support systems that seek to subvert these norms [69, 150, 286]. These highly problematic societal norms are founded in the historical oppression of various groups of people, and manifest today in many ways including economic disadvantages, stereotypes about who can succeed in certain career paths, and racist and/or sexist barriers to opportunity, including educational advancement. While these norms are, by definition, specific to a particular culture or even country, they are nonetheless pervasive and oppressive and demand attention to rectify these historical wrongs.

Much important work has been done on building critical race and/or gender theories of STEM education [136, 137, 252, 160, 14, 93, 213, 261, 107, 198, 245, 242]. In one study, Bancroft (2018) lays out a "critical capital theory," using varying forms of capital (economic, social, and cultural) to examine persistence through graduation in STEM doctoral programs and to contextualize the mechanisms behind racial inequities in STEM education [7]. The idea that race, gender, or another demographic characteristic alone cannot fully explain the intricacies of the obstacles that students face is rooted in the framework of intersectionality [68, 64, 193, 54, 197]. In particular, the combination of different aspects of an individual's social identity (e.g., gender, race, first-generation college status, and socioeconomic status) leads to unique levels of disadvantages that cannot be explained by simply adding together the effects of the individual components of identity [68]. For example, according to the framework of intersectionality, in many STEM disciplines where the societal norm expects that students are white men, the experience of a black woman is not a simple sum of the experiences of white women and black men [54, 197].

With an eye toward this intersectional approach to critical theory, we seek to understand the relationship between four different aspects of student identity that can lead to obstacles in STEM education: race/ethnicity, gender, low-income status, and first-generation college student status. The students disadvantaged by low-income or first-generation status are likely to experience a lack of resources relative to their more privileged peers [156, 81, 149]. Women and underrepresented minority students are susceptible to additional stress and anxiety from stereotype threat (i.e., the fear of confirming stereotypes pertaining to their identity) which is not experienced by their majority group peers [160, 136, 107, 198, 245, 4, 70, 95, 94, 31, 46, 29, 15, 63, 126]. In summary, the different mechanisms by which students belonging to each demographic characteristic can be disadvantaged are as follows.

- Race/Ethnicity: Students belonging to underrepresented minority (URM) groups may experience stereotype threat that causes anxiety and robs the students of their cognitive resources, particularly during high-stakes testing.
- Gender: There are pervasive societal biases against women succeeding in many STEM disciplines which can result in stereotype threat.
- Low-Income Status: Low-Income (LI) students are more likely to need to work to support themselves, reducing their time and energy available to devote to their studies, in addition

to anxiety due to the financial burden of attending college. These burdens are in addition to other factors that low-income students may be more likely to face, such as lower quality preparation for college.

- First-Generation Status: First-Generation (FG) students may lack the resources of encouragement, advice, and support that are available more readily to students with degree-holding parents. This lack of resources can make FG students more susceptible to the stress of the unknown in college.

All of these mechanisms can produce an inequitable learning environment wherein students belonging to any of these groups are forced to work against obstacles that their peers do not have. The framework of intersectionality asserts that for students that belong to more than one of these groups, complex interactions between these different obstacles can result in compounded disadvantages that are not a simple sum of the individual effects [68, 64, 193, 54, 197].

In order to measure the long-term effects of these systemic disadvantages, we will investigate the academic achievement of students belonging to these various demographic groups over the course of their studies at one large public research university using 10 years of institutional data. By grouping students according to their demographic background, we will be able to investigate how different combinations of obstacles affect student grade point averages.

7.1.1 Research Questions

Our research questions regarding the intersectional relationships between demographic characteristics and academic achievement are as follows.

RQ1. Are there differences in the overall or STEM grades earned by students belonging to different demographic groups (i.e., underrepresented minority, low-income status, and first-generation college student status)?

RQ2. Do any patterns observed in RQ1 differ for men and women?

RQ3. Do grades earned in STEM courses alone exhibit similar demographic patterns as grades earned in all courses?

RQ4. What are the trends over time in the mean GPA of these different demographic groups among different clusters of majors (i.e., computer science, engineering, mathematics, and physical science majors, other STEM majors, and non-STEM majors)?

7.2 Methodology

7.2.1 Sample

Using the Carnegie classification system, the university at which this study was conducted is a public, high-research doctoral university, with balanced arts and sciences and professional schools, and a large, primarily residential undergraduate population that is full-time and reasonably selective with low transfer-in from other institutions [133].

The university provided for analysis the de-identified institutional data records of students with Institutional Review Board approval. In this study, we examined these records for $N = 24,567$ undergraduate students enrolled in three colleges within the university: the colleges of Arts and Sciences, Computing and Information, and Engineering. This sample of students includes all of those from ten cohorts who met several selection criteria, namely that the student had first enrolled at the university in a Fall semester from Fall 2005 to Fall 2014, inclusive, and the institutional data on the student was not missing or unspecified for any of the following measures: gender, race/ethnicity, parental education level, and family income. This sample of students is 50% female and had the following race/ethnicities: 79% White, 9% Asian, 7% Black, 3% Hispanic, and 2% other or multiracial. Further, this sample is 16% first-generation college students and 21% "low-income" students (to be defined in the following section).

We acknowledge that gender is not a binary construct, however in self-reporting their gender to the university students were given the options of "male" or "female" and so those are the two self-reported genders that we are able to analyze. There were 39 students who had met all other selection criteria but who had not indicated any gender on the survey, these students were removed from the sample and are not included in the reported sample size or any analyses.

7.2.2 Measures

7.2.2.1 Demographic Characteristics Four primary measures are the demographic characteristics mentioned in the previous section, namely gender, race/ethnicity, parental education level, and family income. All of these were converted into binary categories intended to distinguish between the most and least privileged students on each measure.

- *Gender.* Gender was reported as a binary category to begin with (either "male" or "female"), therefore no further steps were required.
- *First-generation.* Students for whom both parents had a highest completed level of education of high school or lower were grouped together as "first-generation" (FG) college students and correspondingly students for whom at least one parent had earned a college degree were labeled non-FG.
- *Low-income.* Students whose reported family Adjusted Gross Income (AGI) was at or below 200% of the federal U.S. poverty line were categorized as "low-income" (LI), and those above 200% of the poverty line as non-LI [52, 134].
- *Underrepresented minority.* All students who identified as any race or ethnicity other than White or Asian were grouped together as "underrepresented minority" (URM) students, including multiracial students who selected White and/or Asian in addition to another demographic option. Students who only identified as White and/or Asian students were categorized as non-URM students.

7.2.2.2 Academic Performance Measures of student academic performance were also included in the provided data. High school GPA was provided by the university on a weighted scale from 0-5 that includes adjustments to the standard 0-4 scale for Advanced Placement and International Baccalaureate courses. The data also include the grade points earned by students in each course taken at the university. Grade points are on a 0-4 scale with A = 4, B = 3, C = 2, D = 1, F = 0, where the suffixes "+" and "−" add or subtract, respectively, 0.25 grade points (e.g. B− = 2.75), with the exception of A+ which is reported as the maximum 4 grade points. The courses were categorized as either STEM or non-STEM courses, with STEM courses being those courses taken from any of the following departments: biological sciences, chemistry, computer science, economics, any engineering department, geology and environmental science, mathematics, neuroscience, physics and astronomy, and statistics. We note that for the purposes of this paper, "STEM" does not include the social sciences other than economics, which has been included due to its mathematics-intensive content.

7.2.2.3 Year of Study Finally, the year in which the students took each course was calculated from the students' starting term and the term in which the course was taken. Since the sample only includes students who started in fall semesters, each "year" contains courses taken in the fall and subsequent spring semesters, with courses taken over the summer omitted from this analysis. For example, if a student first enrolled in Fall 2007, then their "first year" occurred during Fall 2007 and Spring 2008, their "second year" during Fall 2008 and Spring 2009, and so on in that fashion. If a student is missing both a fall and spring semester during a given year but subsequently returns to the university, the numbering of those post-hiatus years is reduced accordingly. If instead a student is only missing one semester during a given year, no corrections are made to the year numbering. In this study we consider up through the students' sixth year of study or the end of their enrollment at the studied institution, whichever comes first.

7.2.3 Analysis

The primary method by which we grouped students in this analysis was by their set of binary demographic categories. This grouping was performed in two different ways. First, use of all four binary categories (gender, FG, LI, URM) resulted in sixteen mutually exclusive groups (e.g., "female, FG+URM" or "male, LI"). Second, use of all categories except gender resulted in eight mutually exclusive categories.

We calculated each student's yearly (i.e., not cumulative) grade point average (GPA) across courses taken in each year of study from the first to sixth years. In addition, we calculated the student's yearly STEM GPA, that is, the GPA in STEM courses alone. Then, using the aforementioned grouping schemes, we computed the mean GPA in each demographic group as well as the standard error of the mean separately for each year of study [98]. Further, in the case of grouping by gender, we computed the effect size of the gender differences within each demographic group using Cohen's d, which is typically interpreted using minimum cutoff values for "small" ($d = 0.20$), "medium" ($d = 0.50$), and "large" ($d = 0.80$) effect sizes [65, 201, 195].

All analyses were conducted using R [226], making use of the package `tidyverse` [279] for data manipulation and plotting.

7.3 Results

7.3.1 GPA Trends by Demographic Group: "Dinosaur Plots"

In order to answer **RQ1**, we plotted in Fig. 14 the mean GPA earned by students in each demographic group, including gender as a grouping characteristic. We start with overall GPA, rather than STEM GPA alone, in order to provide context for the results in STEM GPA and identify trends that may or may not be present when viewing STEM grades alone. Groups are ordered from left to right first by the ascending number of selected characteristics and then alphabetically. Mean GPA is plotted separately (i.e., not cumulatively) for each year of study from the first to sixth year. Setting aside the gender differences for a moment, we note that the general GPA trends by demographic group in Fig. 14 follow a shape resembling the neck, back, and tail of a sauropod, and so accordingly we refer to the plots in Fig. 14 as "dinosaur plots." This shape is clearest in the plots for the first through fourth years, as the sample size drops significantly in the fifth year as the majority of students graduate.

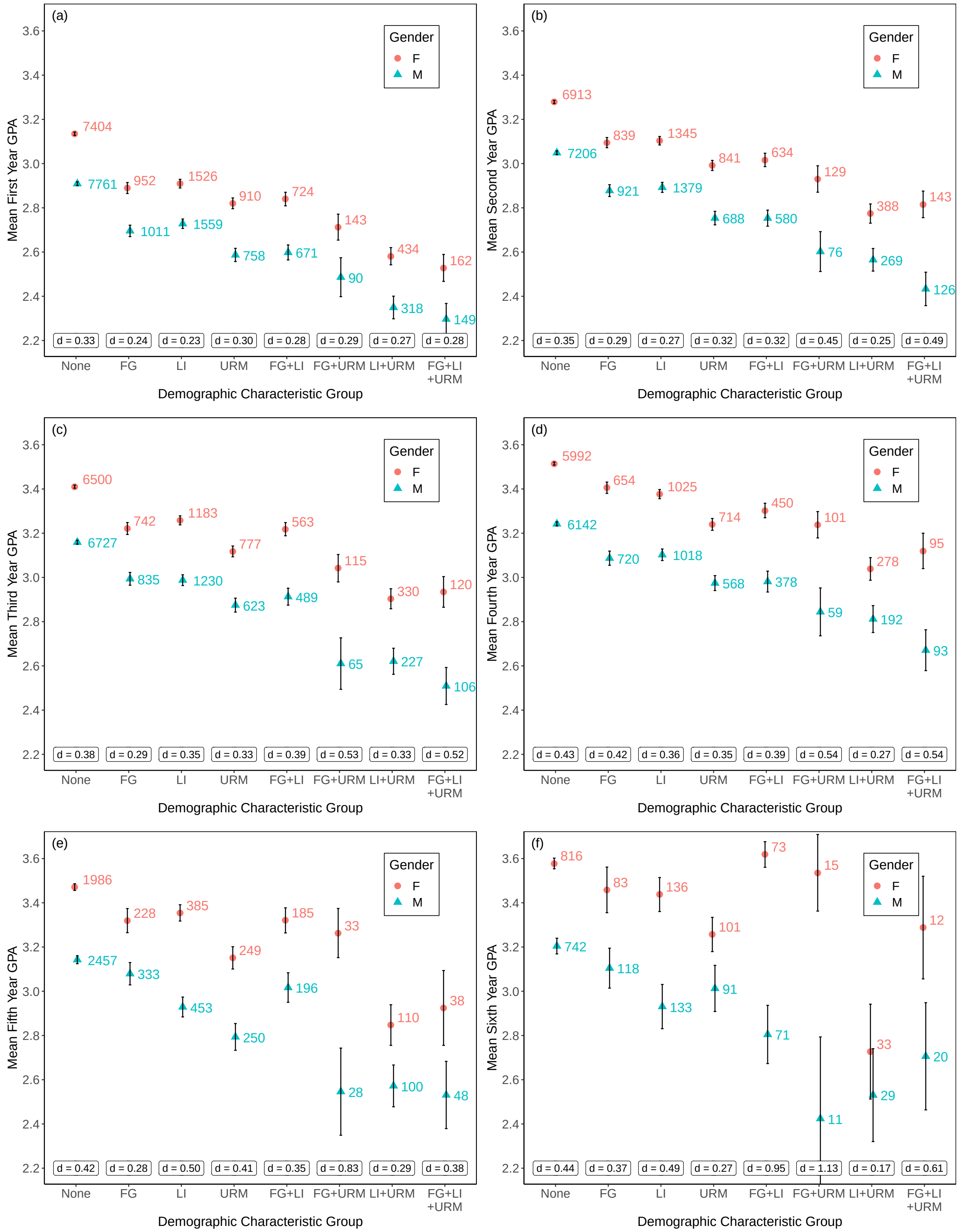
Looking more closely at Fig. 14, particularly the first four years, we see that the "neck" is consistently comprised of the group of students with the most privileges, namely those students that are non-FG, non-LI, and non-URM. Following this, the "back" is relatively flat across the next four groups, namely students that are FG only, LI only, URM only, or FG and LI. Notably, the URM group of students typically have the lowest mean GPA within this set of demographic groups. Finally, the "tail" consists of the final three groups, FG+URM, LI+URM, and FG+LI+URM. The mean GPA in this set of groups tends to decrease from left to right in the plots. Notably, the four groups that contain URM students are consistently in the lowest four or five mean GPAs.

7.3.2 Intersectionality with Gender

We now turn our attention to the differences between men and women in Fig. 14 in order to answer **RQ2**. We note in particular that across all demographic groups women's mean GPA is roughly 0.2 grade points higher than men's. The effect sizes (Cohen's d) of this difference range from small to medium [65]. This difference in mean GPA earned is substantial enough to indicate a change in letter grade, given that the grading system at the studied university uses increments of 0.25 grade points for letter grades containing "+" or "−." Further, this trend holds in the fifth year (Fig. 14e) and sixth year (Fig. 14f), with some exceptions in demographic groups with particularly low sample sizes after the fourth year.

7.3.3 STEM GPA Trends

In order to answer **RQ3**, Figure 15 plots students' mean STEM GPA in a similar manner to Fig. 14. We note that the general "dinosaur" pattern discussed in Fig. 14 also holds at least for the first and second years (Figs. 15a and 15b, respectively). In the third year and beyond, the general features of the trend continue to hold, with the most privileged students having the highest mean GPA, followed by those with one disadvantage as well as the first-generation and low-income group, followed by the remaining groups of URM students with one or more additional disadvantages. However, in these later years, the finer details of the plots noted before fall away in favor of a sharper mean GPA decrease for URM students with at least one additional disadvantage in the third year (Fig. 15c) and a more gradual decrease across all groups in the fourth year (Fig. 15d) and fifth year (Fig. 15e). When restricting



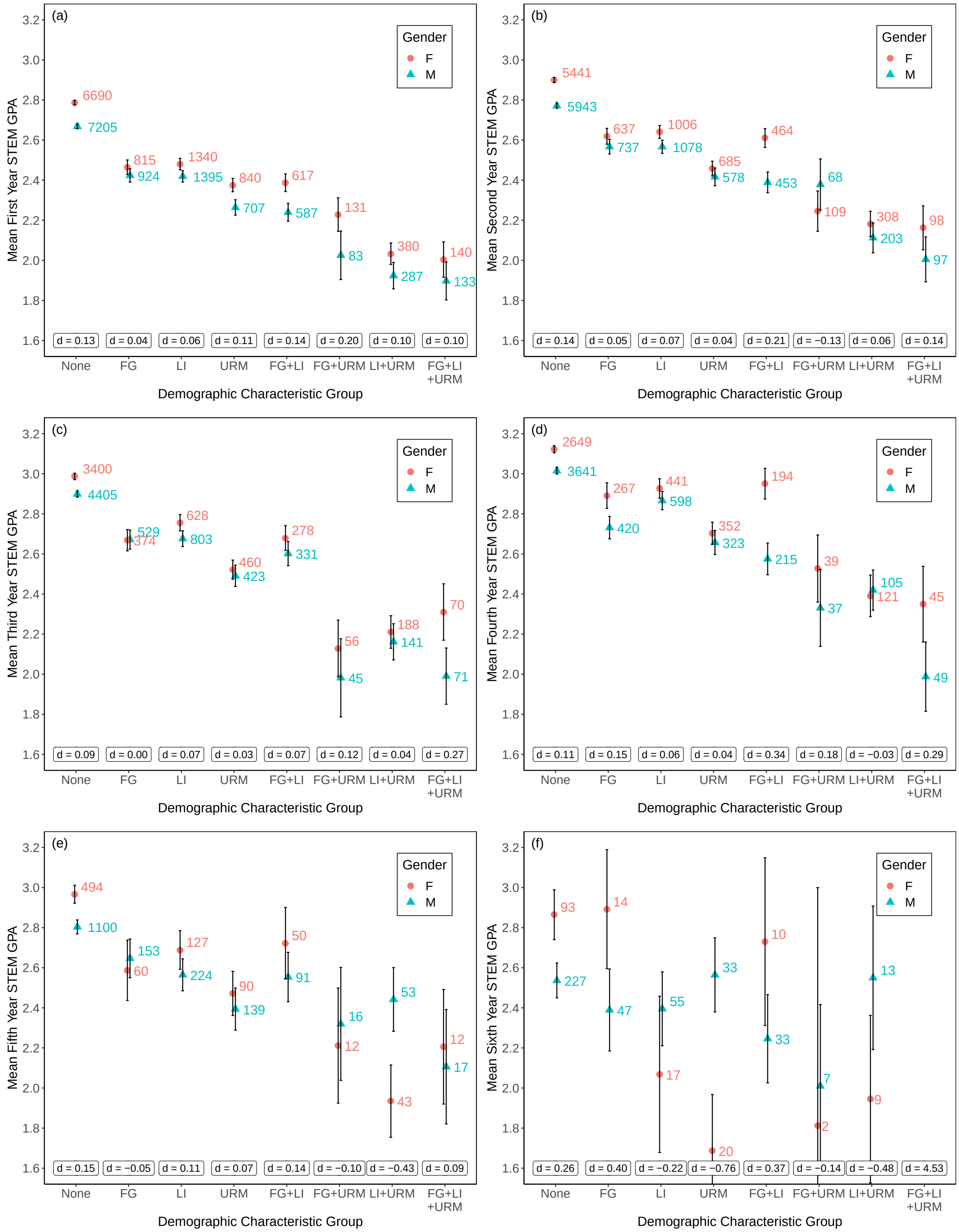
(Caption on next page.)

Figure 14: (Previous page.) Average GPA of each demographic group. Students are binned into separate demographic groups based on their status as first-generation (FG), low-income (LI), and/or underrepresented minority (URM) students. The men and women in each demographic group are plotted separately. The mean GPA in all courses taken by students in each demographic group is plotted along with the standard error on the mean, with a separate plot for each of the (a) first, (b) second, (c) third, (d) fourth, (e) fifth, and (f) sixth years. The sample size is reported by each point, and Cohen's d [65] measuring the effect size of the gender difference in each group is reported.

the GPA calculations to STEM courses, the sample size becomes too small in the sixth year (Fig. 15f) to draw meaningful conclusions.

We further observe a trend of students earning higher grades on average in later years, although the rise from the first to the fourth year is somewhat lower in STEM GPA than in overall GPA. Notably, while in overall GPA this trend seemed to be somewhat universal across demographic groups, in Fig. 15 we see a quicker rise in mean STEM GPA over time for the more privileged students than the less privileged students, particularly comparing the leftmost and rightmost groups.

Regarding gender differences, Fig. 15 shows smaller gender differences in STEM GPA than those observed in overall GPA in Fig. 14. While in overall GPA women earned roughly 0.2 grade points more than men on average, in STEM GPA that difference is much less consistent and typically ranges from 0 to 0.1 grade points. For many demographic groups we see no significant differences between men and women's mean STEM GPA. We do see that there is still a consistent STEM GPA gender difference, albeit smaller than in Fig. 14, among the group of the most privileged students (i.e., those with "None" of the disadvantages). There is also a STEM GPA gender difference among first-generation low-income but non-URM students, however this difference is less consistent and in fact briefly vanishes in the third year.



(Caption on next page.)

Figure 15: (Previous page.) Average STEM GPA of each demographic group. Students are binned into separate demographic groups based on their status as first-generation (FG), low-income (LI), and/or underrepresented minority (URM) students. The men and women in each demographic group are plotted separately. The mean GPA in all courses taken by students in each demographic group is plotted along with the standard error on the mean, with a separate plot for each of the (a) first, (b) second, (c) third, and (d) fourth, (e) fifth, and (f) sixth years. The sample size is reported by each point, and Cohen's d [65] measuring the effect size of the gender difference in each group is reported.

7.3.4 GPA Trends By Major Over Time

In order to better understand the trends over time in both overall and STEM GPA and answer **RQ4**, we plotted the mean GPA by year in Fig. 16 and mean STEM GPA by year in Fig. 17. In these plots, we have not separated men and women and instead focus on the other demographic characteristics while further grouping students into three different groups of majors in order to understand if these trends differ for students in different areas of study. Further, since the sample size becomes quite small in years five and six for many of the demographic groups of interest, we plot only the mean GPA over the first four years. In Figs. 16a and 17a, we plot the mean overall and STEM GPA, respectively, of all students. In the other subfigures, we plot the mean GPA earned by students majoring in different clusters of majors. In particular, we plot the mean GPA of engineering (including computer science), mathematics, and physical science (i.e., chemistry and physics) majors in Figs. 16b and 17b, the remaining STEM majors in Figs. 16c and 17c, and non-STEM majors in Figs. 16d and 17d.

These plots make clearer some of the trends noted earlier, especially the rise in mean GPA over time from the first to the fourth year. However, we can now see that this is not universally true since the first-generation URM students have a drop in mean GPA in the second year for physical science majors (Fig. 16b), and in the third year for other STEM

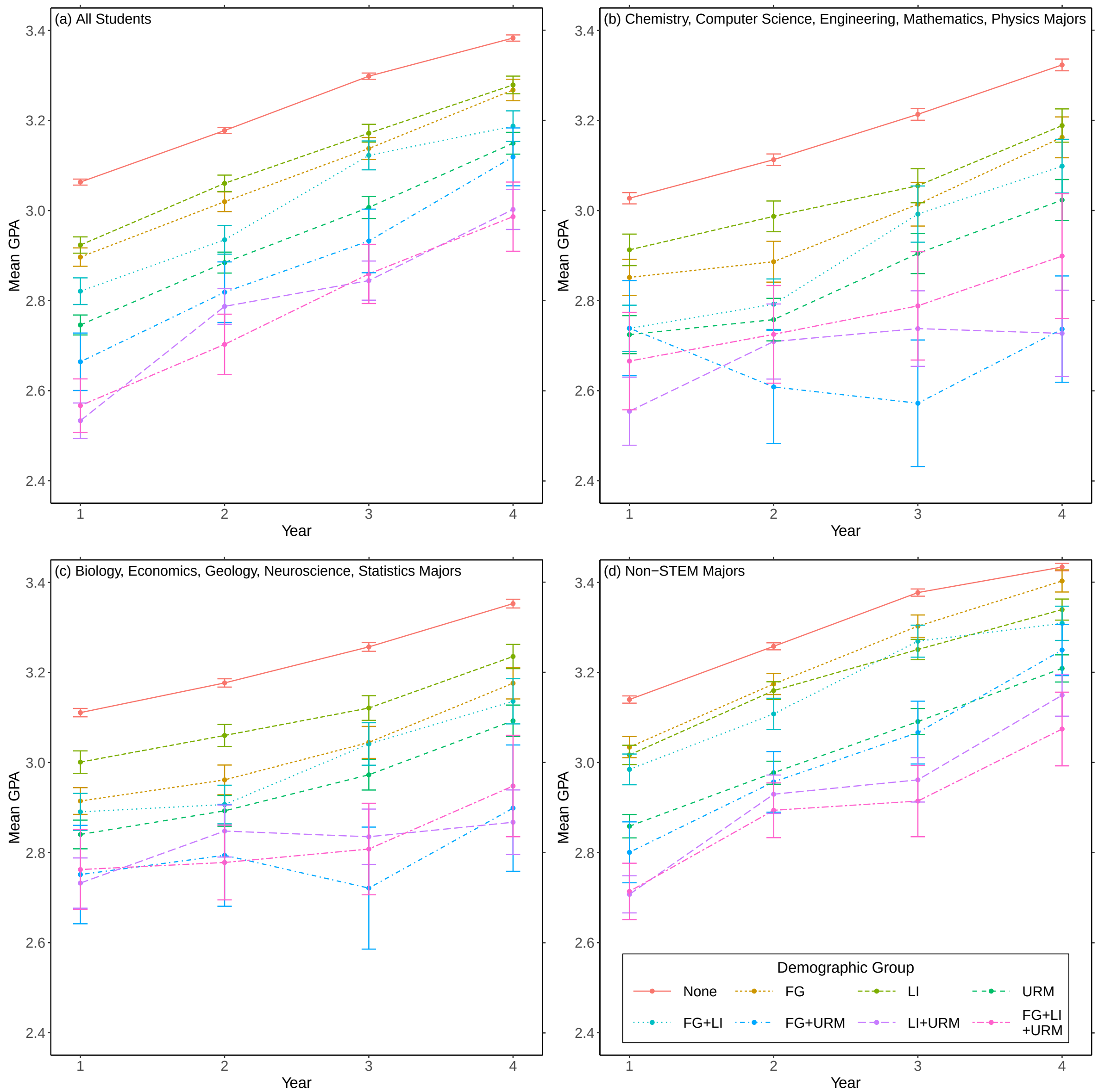


Figure 16: Students are binned into separate demographic groups as in Fig. 14, but not separated by gender. The mean GPA in all courses of each group is plotted over time from year one to four, along with the standard error of the mean. The plots show this for four subpopulations: (a) all students; (b) chemistry, computer science, engineering, mathematics, and physics students; (c) biology, economics, geology, neuroscience, and statistics students; and (d) non-STEM students including psychology.

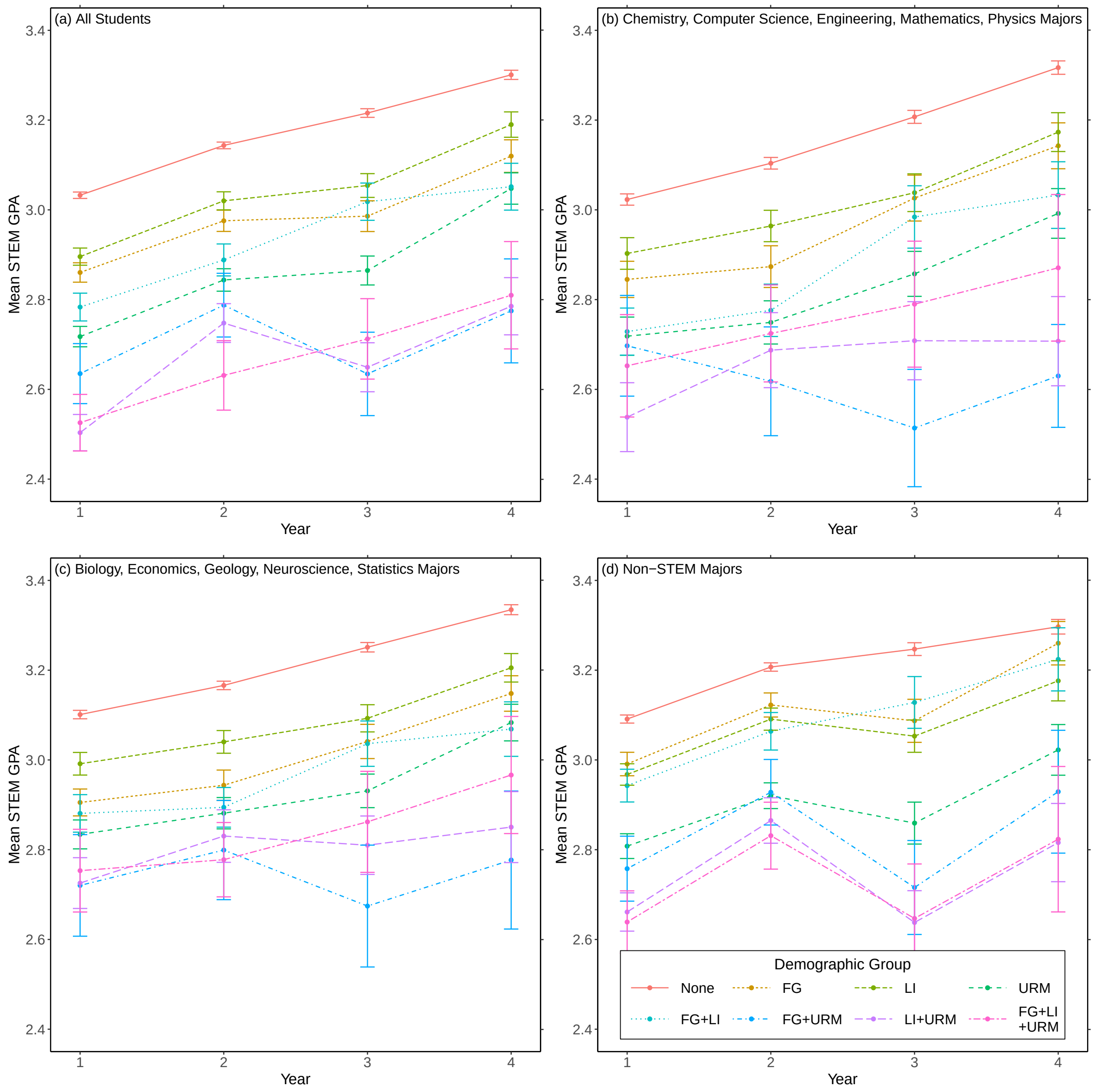


Figure 17: Students are binned into separate demographic groups as in Fig. 15, but not separated by gender. The mean GPA in STEM courses of each group is plotted over time from year one to four along with the standard error of the mean. The plots show this for four subpopulations: (a) all students; (b) chemistry, computer science, engineering, mathematics, and physics students; (c) biology economics, geology, neuroscience, and statistics students; and (d) non-STEM students including psychology.

majors (Fig. 16c). This trend is even more noticeable in STEM GPA (Fig. 17), where the mean STEM GPA of the group of first-generation URM students drops in the third year for every subpopulation by major.

7.4 Discussion

To start, we consider how much the current system disadvantages students who are first-generation, low-income, or underrepresented minority but not a combination of the two. Discussing these groups first is helpful in setting the stage for a more complex discussion of the intersectionality of these various demographic characteristics. We find in Figs. 14 and 15 that not all of these disadvantages are equal. In particular, non-URM students who have one disadvantage, namely the first-generation (but not low-income) and low-income (but not first-generation) students, still earn slightly higher grades than even the URM students who are not low-income or first-generation. Notably, this trend (the "back" of the dinosaur plots) is similar in both overall grades (Fig. 14) and in STEM grades alone (Fig. 15). The size of this mean grade difference varies from year to year, but in STEM grades it can reach as high as about 0.25 grade points, which at the studied institution is the difference between, for example, a B and B+ or B− grade.

The group with the grades most similar to these non-first-generation, non-low-income URM students are the first-generation, low-income non-URM students, who earn both overall (Fig. 14) and STEM (Fig. 15) grades similar to or very slightly higher than the URM students. One explanation could be that the lack of resources available due to being first-generation or low-income is not as severe an obstacle as the stereotype threat experienced by URM students.

Turning then to the "tail" in the dinosaur plots, we find that consistently the most disadvantaged students in both overall grades (Fig. 14) and STEM grades (Fig. 15) are the URM students with at least one additional obstacle. In this case, it appears that the intersection of being low-income and URM is the most disadvantageous combination, with no notable difference in either Fig. 14 or Fig. 15 among these students whether or not they are also first-generation. Meanwhile, the first-generation URM students who are not low-income sometimes have a slightly higher mean GPA than the low-income URM students (Fig. 14).

Another avenue to investigate intersectionality is how gender interacts with the other demographic groups. Interestingly, in overall GPA (Fig. 14), gender appears to have about the same effect across all demographic groups. That is, there does not appear to be an intersectional effect of gender identity with other identities as measured by overall GPA. However, Fig. 15 shows that this is a context-dependent effect, with the gender gap substantially and unevenly reduced across all groups in mean STEM GPA. For most demographic groups in Fig. 15, the higher overall GPA earned by women in Fig. 14 has vanished completely in STEM GPA. This is consistent with stereotype threat being the mechanism of disadvantage for women, where stereotypes surrounding STEM disciplines unfairly cause stress and anxiety for women [4, 70, 95, 94, 46, 15, 63, 126]. Notably, while the gender gap is reduced nearly to zero for most groups in Fig. 15, there does remain a small consistent gender gap favoring women in the most privileged group of students. In other groups the gender gap in Fig. 15 is inconsistent across years. One explanation could be that the wealth of resources available to them may help to alleviate the stereotype threat.

Taking a more temporal view of these GPA trends, Fig. 16 (overall GPA) and Fig. 17 (STEM GPA) have grouped men and women together in order to focus on the other demographic characteristics more closely. In these plots, the most noteworthy trend is again that, with the sole exception of the first year in Fig. 16b, the four groups with the lowest mean GPA (Fig. 16) and STEM GPA (Fig. 17) across the first four years are always the four groups containing URM students. Notably, this trend is true regardless of which group of majors we investigate. The consistency of this result is particularly striking, showing that the most otherwise disadvantaged non-URM students have fewer obstacles to success than even the most privileged URM students among all students.

Focusing further on the STEM GPA of STEM majors in Figs. 17b and 17c, we see that while non-URM students consistently rise in mean GPA over time, the same is not true for all URM students. In particular, the first-generation URM students who major in chemistry, computer science, engineering, mathematics, or physics (Fig. 17b) experience a steady decline in mean STEM GPA from year one to two and year two to three. While the standard error of those means is quite large due to a relatively small sample size, that lack of representation for these students could itself be what is hindering their coursework by causing a stereotype threat.

Based upon the frameworks of critical theory and intersectionality, the main implication of these findings is that many students who come from less privileged backgrounds are not being adequately supported in college in order to catch up with the privileged students [69, 150, 286, 110, 258, 261, 242, 136, 137, 68, 64, 193, 54, 197]. The disadvantages of these less privileged students manifest as lower mean overall and STEM GPA for those demographic groups. In order to promote equity and inclusion, it is crucial that these students are provided appropriate mentoring, guidance, scaffolding, and support in college so that these obstacles can be cleared for students who have been put at a disadvantage relative to their peers through no fault of their own [34]. We note that these demographic groups with more disadvantages are likely to consist of students who had K-12 education from schools with fewer resources and less well-prepared teachers than those of the more privileged students, with high school being an especially important time for disadvantages related to STEM learning increasing [30, 172, 187, 41, 71, 82]. Analyses such as those discussed here can help inform the allocation of resources to support these students, with efforts to reduce the classroom stereotype threat of URM students and creating a low-anxiety environment in which all students have a high sense of belonging and can participate fully without fear of being judged being clear priorities. Additional resources to support low-income and/or first-generation students, e.g., financial support and timely advising pertaining to various academic and co-curricular opportunities, are also important in order to level the playing field and work towards a goal of all students succeeding in college, regardless of their race/ethnicity, socioeconomic status, and parental education history.

7.5 Acknowledgments

This research is supported by the National Science Foundation Grant DUE-1524575 and the Sloan Foundation Grant G-2018-11183.

8.0 Inequity in Grades and Persistence: Gender

8.1 Introduction and Theoretical Framework

Increasingly, Science, Technology, Engineering, and Mathematics (STEM) departments across the US are focusing on using evidence to improve the learning of all students regardless of their background and making learning environments equitable and inclusive [136, 137, 188, 151, 172, 171, 186, 39, 37, 38, 119, 72, 118]. However, women are still severely underrepresented in many STEM disciplines [200, 199]. In order to understand the successes and shortcomings of the current state of education, the use of institutional data to investigate past and current trends is crucial. In the past few decades, institutions have been keeping increasingly large digital databases of student records. We have now reached the point where there are sufficient data available at many institutions for robust statistical analyses using data analytics that can provide invaluable information for transforming learning environments and making them more equitable and inclusive for all students [6, 217]. Studies utilizing many years of institutional data can lead to analyses that were previously limited by statistical power. This is particularly true for studies of performance and persistence in STEM programs that rely on large sample sizes [212, 166, 92, 172, 192, 167, 211, 181, 283, 151, 236, 171, 186].

In this study, we use 10 year institutional data from a large state-related research university to investigate how patterns of student major-declaration and subsequent degree-earning may differ for men and women. The theoretical framework for this study has two main foundations: critical theory and expectancy value theory.

Critical theories, e.g., of race and gender, focus on historical sources of inequities within society, that is, societal norms that perpetuate obstacles to the success of certain groups of disadvantaged people [106, 69, 150, 286, 110, 258, 261, 242, 188]. Critical theory tells us that the dominant group in a society perpetuates these norms, which are born out of their interests, and pushes back against support systems that seek to subvert these norms [69, 150, 286]. In our case, critical gender theory provides a historical perspective on the much-studied gender inequities in STEM.

Much important work has been done that relates to critical theory of gender in STEM education [136, 137, 14, 93, 213, 261, 242, 166, 243, 44, 45, 188, 99]. One mechanism by which historical societal stereotypes and biases about gender can influence student choice of major is proposed by Leslie *et al.*, who showed that disciplines with a higher attribution of "brilliance" also have a lower representation of women [158] due to pervasive stereotypes about men being "brilliant" in those disciplines. These brilliance-attributions affect all levels of STEM education, starting with early childhood where girls have already acquired these notions that girls are not as brilliant as boys [26, 25], which can later influence their interest in pursuing certain STEM disciplines [28], and even affect how likely they are to be referred for employment in these disciplines in professional contexts [27].

Eccles's expectancy value theory (EVT) [88, 87, 86] is another framework that is central to our investigation. EVT states that a student's persistence and engagement in a discipline are related to the student's expectancy about their success as well as how the student values the task [88, 87, 86]. In an academic context, "expectancy," which refers to the individual's beliefs about their success in the discipline, is closely related to Bandura's construct of self-efficacy, defined as one's belief in one's capability to succeed at a particular task or subject [8, 9, 10, 11, 12, 13, 88, 87, 86].

There are four main factors that influence students' expectancy or self-efficacy, namely vicarious experiences (e.g., instructors or peers as role models), social persuasion (e.g., explicit mentoring, guidance, and support), level of anxiety [8, 9, 10, 11, 12, 13], and performance feedback (e.g., via grades on assessment tasks). Women generally have lower self-efficacy than men in many STEM disciplines because these four factors negatively influence them [136, 137, 14, 93, 213, 261, 242]. For example, in many STEM fields women are underrepresented in their classrooms, and less likely to have a female role model among the faculty [200, 199]. Further, the stereotypes surrounding women in many STEM disciplines can affect how they are treated by mentors, even if such an effect is subconscious [4, 136, 95, 94, 63, 31, 46, 126]. Moreover, women are susceptible to stress and anxiety from stereotype threat (i.e., the fear of confirming stereotypes about women in many STEM disciplines) which is not experienced by their male peers [4, 136, 95, 94, 63, 31, 46, 126]. This stress and anxiety can rob them of their cognitive resources, especially during high-stakes assessments such as exams.

Expectancy can influence grades earned as well as the likelihood to persist in a program [8, 9, 10, 11, 12, 13]. Stereotype threats that women in many STEM disciplines experience can increase anxiety in learning and test-taking situations and lead to deteriorated performance. Since anxiety can increase when performance deteriorates, these factors working against women in STEM can force them into a feedback loop and hinder their performance further, which can further lower their self-efficacy and can continue to affect future performance [8, 9, 10, 11, 12, 13].

In EVT, value is typically defined as having four facets: intrinsic value (i.e., interest in the task), attainment value (i.e., the importance of the task for the student's identity), utility value (i.e., the value of the task for future goals such as career), and cost (i.e., opportunity cost or psychological effects such as stress and anxiety) [88, 87, 86]. In the context of women's enrollment and persistence in many STEM disciplines, the societal stereotypes can influence all facets of the students' value of these STEM disciplines. Intrinsic value can be informed by societal stereotypes and brilliance-attributions of the STEM disciplines, and attainment and utility values can be further tempered by these stereotypes. Utility value is an important facet of student education in STEM, since a degree in a STEM field provides many job opportunities for graduating students. In addition, the psychological cost of majoring in these disciplines can be inflated by the stereotype threat. All of these effects can conspire to suppress the likelihood of women choosing and/or persisting in various STEM disciplines.

In order to measure the long-term effects of these systemic disadvantages, we investigate the differences in attrition rates and choices of major of men and women over the course of their studies at one large public research university using 10 years of institutional data. Since these disadvantages to students can be context-dependent, we will consider the attrition rates in many different STEM majors and non-STEM majors in order to understand the trends in each discipline.

8.1.1 Research Questions

Our research questions regarding the relationships between gender, degree attainment, attrition, performance and persistence pertaining to a college degree over a 10 year period are as follows.

RQ1. How many students major in each discipline? How many men and women major in each discipline?

RQ2. Do rates of attrition from the various majors differ? Do rates of attrition from the various majors differ for men and women?

RQ3. Among those students who drop a given major, what degree, if any, do those students earn? How do these trends differ for men and women?

RQ4. What fraction of declared majors ultimately earn a degree in that major in each STEM subject area? How do these trends differ for men and women?

RQ5. What are the GPA trends over time among students who earn a degree in a given major and those who drop that major? How do these trends differ for men and women?

8.2 Methodology

8.2.1 Sample

Using the Carnegie classification system [133], the university at which this study was conducted is a public, high-research doctoral university, with balanced arts and sciences and professional schools, and a large, primarily residential undergraduate population that is full-time and reasonably selective with low transfer-in from other institutions.

The university provided for analysis the de-identified institutional data records of students with Institutional Review Board approval. In this study, we examined these records for $N = 18,319$ undergraduate students enrolled in two schools within the university: the School of Engineering and the School of Arts and Sciences. This sample of students includes all of those from ten cohorts who met several selection criteria, namely that the students had first enrolled at the university in a Fall semester, had provided the university with a self-reported gender, and the students had either graduated and earned a degree, or had not attended the university for at least a year as of Spring 2019. This sample of students is 49.9% female and had the following races/ethnicities: 77.7% White, 11.1% Asian, 6.8% Black, 2.5% Hispanic, and 2.0% other or multiracial.

8.2.2 Measures

8.2.2.1 Gender In this study, we focus on gender differences in student trajectories as they progress towards degrees. We acknowledge that gender is not a binary construct; however, in self-reporting their gender to the university students were given the options of "male" or "female" and so those are the two self-reported genders that we are able to analyze. The student responses to this question were included in the institutional data provided by the university. Very few students opted not to provide a gender, and so were not considered in this study. We used the answers of those students who chose either "male" ("M") or "female" ("F") to group students in order to calculate summary statistics on the measures described in this section.

8.2.2.2 Academic Performance Measures of student academic performance were also included in the provided data. High school GPA was provided by the university on a weighted scale from 0-5 that includes adjustments to the standard 0-4 scale for Advanced Placement and International Baccalaureate courses. The data also include the grade points earned by students in each course taken at the university. Grade points are on a 0-4 scale with A = 4, B = 3, C = 2, D = 1, F = 0, where the suffixes "+" and "−" add or subtract, respectively, 0.25 grade points (e.g., B− = 2.75), with the exception of A+ which is reported as the maximum 4 grade points. The courses were categorized as either STEM or non-STEM courses, with STEM courses being those courses taken from any of the following departments: biological sciences, chemistry, computer science, any engineering department, geology and environmental science, mathematics, neuroscience, physics and astronomy, and statistics. We note that for the purposes of this paper, "STEM" does not include the social sciences.

8.2.2.3 Declared Major and Degree Earned For each student, the data include their declared major(s) in each semester as well as the major(s) in which they earned a degree, if any. The data were transformed into a set of binary flags for each semester, one flag for each possible STEM major as well as specific flags for the non-STEM majors psychology and economics and a general non-STEM category for all other non-STEM majors. A similar set of flags was created for the degrees earned by students. From these flags, we tabulated a number of major-specific measures in each semester, including

- current number of declared majors,
- number of newly declared majors from the previous semester,
- number of dropped majors from the previous semester,
- number of retained majors from the previous semester.

The total number of unique students that ever declared or dropped a major were also computed. The subset of students that dropped each major were further investigated and the major in which they ultimately earned a degree, if any, was determined.

Throughout this paper we group the STEM majors into three clusters: biological sciences (including neuroscience); computer science and engineering; and mathematics (including statistics), chemistry, physics and astronomy, and geology and environmental science (collectively, mathematics and physical science). We additionally consider two non-STEM majors, economics and psychology, separately from the rest of the non-STEM majors. When ordering majors (i.e., in figures and tables), the majors will be presented in the order they are listed in the previous two sentences. Note that "engineering" groups together all engineering majors for departments in the School of Engineering at the studied university. These majors include chemical, computer, civil, electrical, environmental, industrial, and mechanical engineering as well as bioengineering and materials science.

Finally, we will make use of shortened labels for the majors in figures and tables. These shortened labels are defined in Table 17.

8.2.2.4 Year of Study Finally, the year in which the students took each course was calculated from the students' starting term and the term in which the course was taken.

Table 17: A list of the majors considered in this study and the shortened labels used to refer to those majors in tables and figures. Note that "Engineering" is a combination of many engineering majors offered by the School of Engineering.

Major	Short Label
Biological Sciences and Neuroscience	Bio
Computer Science	CS
Engineering	Engr
Mathematics and Statistics	Math
Chemistry	Chem
Physics and Astronomy	Phys
Geology and Environmental Science	Geo
Economics	Econ
Psychology	Psych
Other Non-STEM	Non-STEM

Since the sample only includes students who started in fall semesters, each "year" contains courses taken in the fall and subsequent spring semesters, with courses taken over the summer omitted from this analysis. For example, if a student first enrolled in Fall 2007, then their "first year" occurred during Fall 2007 and Spring 2008, their "second year" during Fall 2008 and Spring 2009, and so on in that fashion. If a student is missing both a fall and spring semester during a given year but subsequently returns to the university, the numbering of those post-hiatus years is reduced accordingly. If instead a student is only missing one semester during a given year, no corrections are made to the year numbering.

8.2.3 Analysis

For each student, we calculate their grade point average (GPA) across courses taken in each year of study from their first to sixth years. In addition, we calculate the student's STEM GPA in each year, that is, the GPA in STEM courses alone. The mean GPA as well as the standard error of the mean are computed for various groupings of students [98].

Further, proportions of students in various groups (i.e., grouped by major and/or gender) are calculated along with the standard error of a proportion [98]. In particular, the proportions we report are

- the proportion of students in each major that are men or women,
- the proportion of men and women, respectively, that declare each subject as a major,
- the proportion of declared majors that drop the major,
- the proportion of those who drop each major that earn a degree in another major, and
- the proportion of all declared majors that ultimately earn a degree in that major.

All analyses were conducted using R [226], making use of the package `tidyverse` [279] for data manipulation and plotting.

8.3 Results

8.3.1 Major Declaration Patterns

There are many angles with which we can approach **RQ1** and investigate patterns of student major declaration. First, Fig. 18 shows the number of students that ever declared each major separately for men and women. These results provide an important context for the upcoming analyses that may be partially explained by the number of students in each major.

Figure 18 begins to hint at gender differences in enrollment patterns, such as a higher proportion of women majoring in non-STEM disciplines than men, or a higher proportion of men majoring in engineering than women. These gender patterns are explored further in Fig. 19 by standardizing the scales in two ways. In Fig. 19a, we consider the populations of each major separately and calculate the percentages of that population that are men or women. This provides insight into what these students might be seeing in the classes for their major. For instance, in the biological sciences there is a roughly even split, so students in biology classes for biology majors might see a classroom that is equally representative of men and women. On the other end of the spectrum, around 80% of both computer science and physics and astronomy majors are men.

Another way to represent the population of these majors is to consider what percentage of all men or women choose each major, as seen in Fig. 19b. While this plot mimics that of Fig. 18, we can now read the differences noted earlier more clearly. In particular, the clearest differences in this view (Fig. 19b) are in engineering (31% of men and 11% of women declare an engineering major), non-STEM (27% of men and 36% of women declare a non-STEM major), and psychology (6% of men and 16% of women declare a psychology major).

Finally, another piece of information about enrollment patterns that is missing from Figs. 18 and 19 is when these students declare each major. Figure 20 shows, for each major, the average term in which students added the major as well as the peak term (that is, the term with the highest number of new students adding the major). This is done separately for all students (Fig. 20a), female students (Fig. 20b), and male students (Fig. 20c).

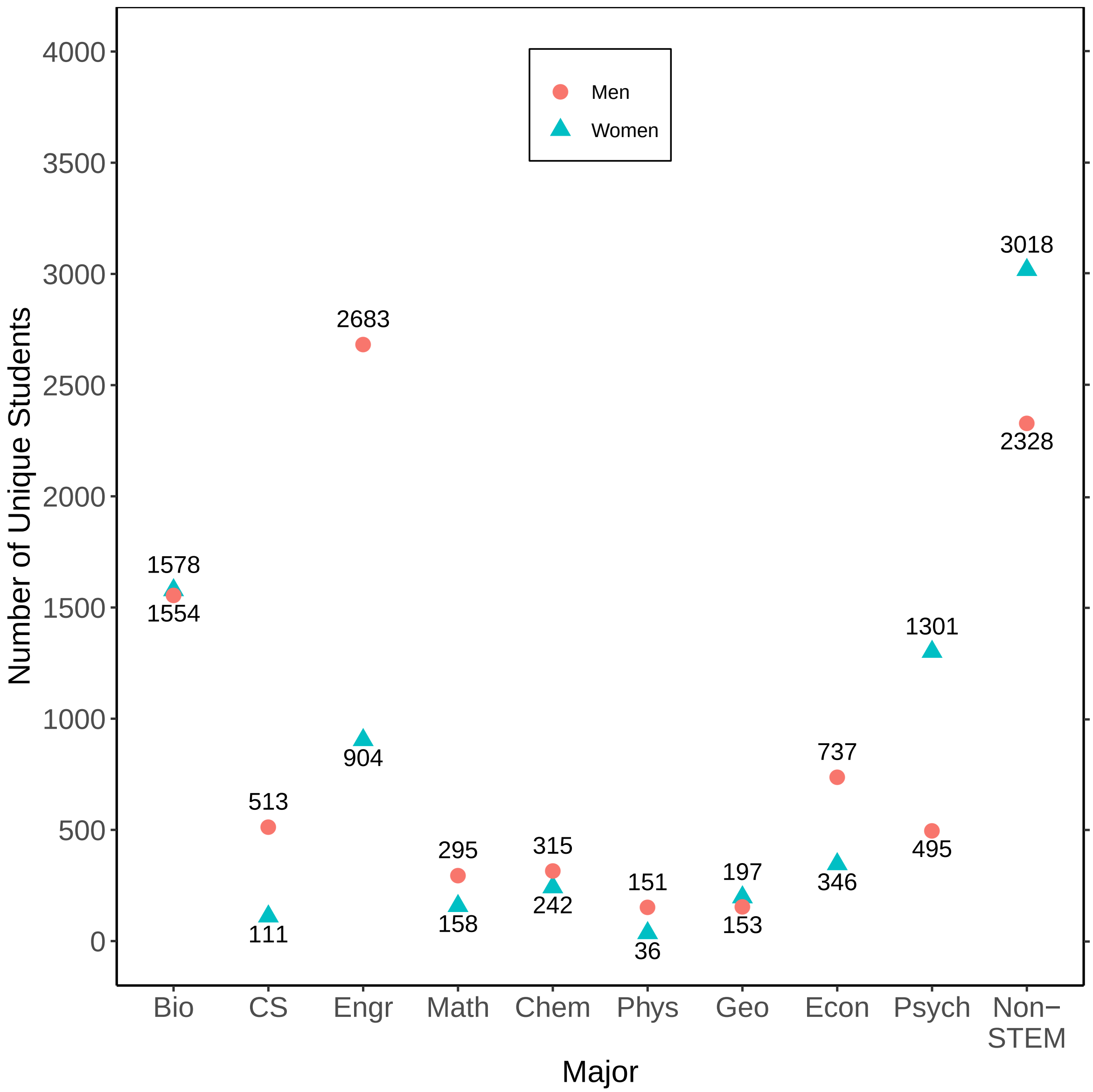


Figure 18: For each major on the horizontal axis, the number of unique students in the sample that ever declared that major is plotted. Since students may change majors or declare multiple majors, some students may contribute to the counts of more than one major. These counts are calculated separately for men and women.

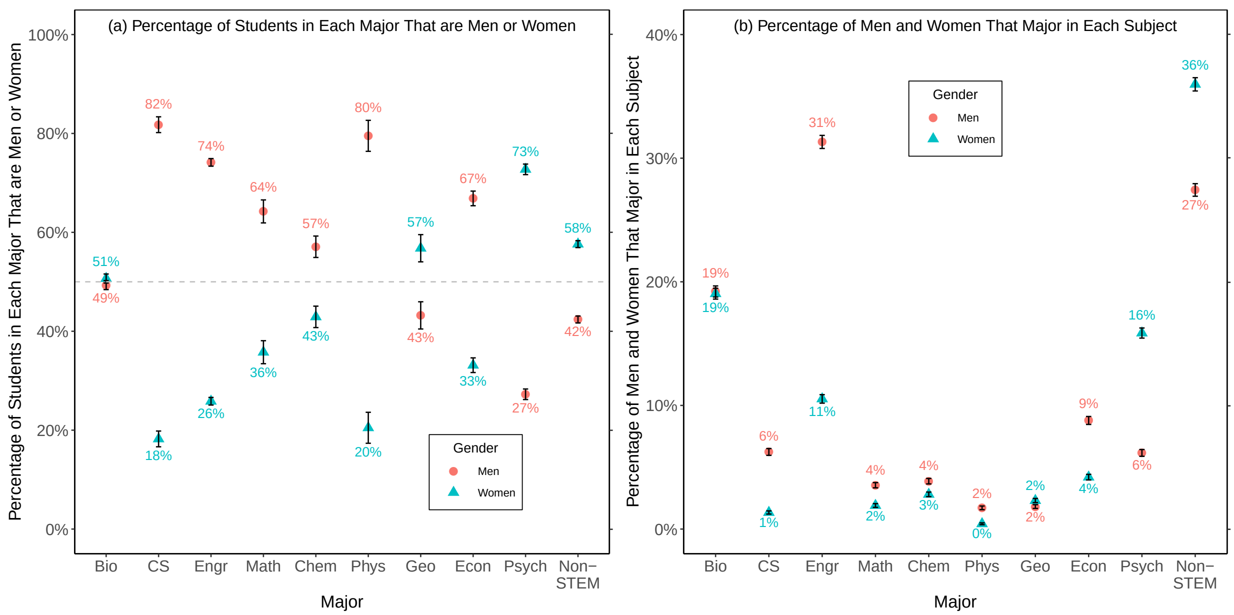


Figure 19: In (a), the percentages of students in each major that are men or women are calculated. A horizontal dashed line of symmetry is shown at 50%. In (b), the percentages of men and women that major in each subject are calculated (i.e., the percentages for each gender group will sum to roughly 100% in this case). Discrepancies in the sum of percentages in (b) may occur due to rounding the listed percentages to the nearest integer as well as students declaring multiple majors.

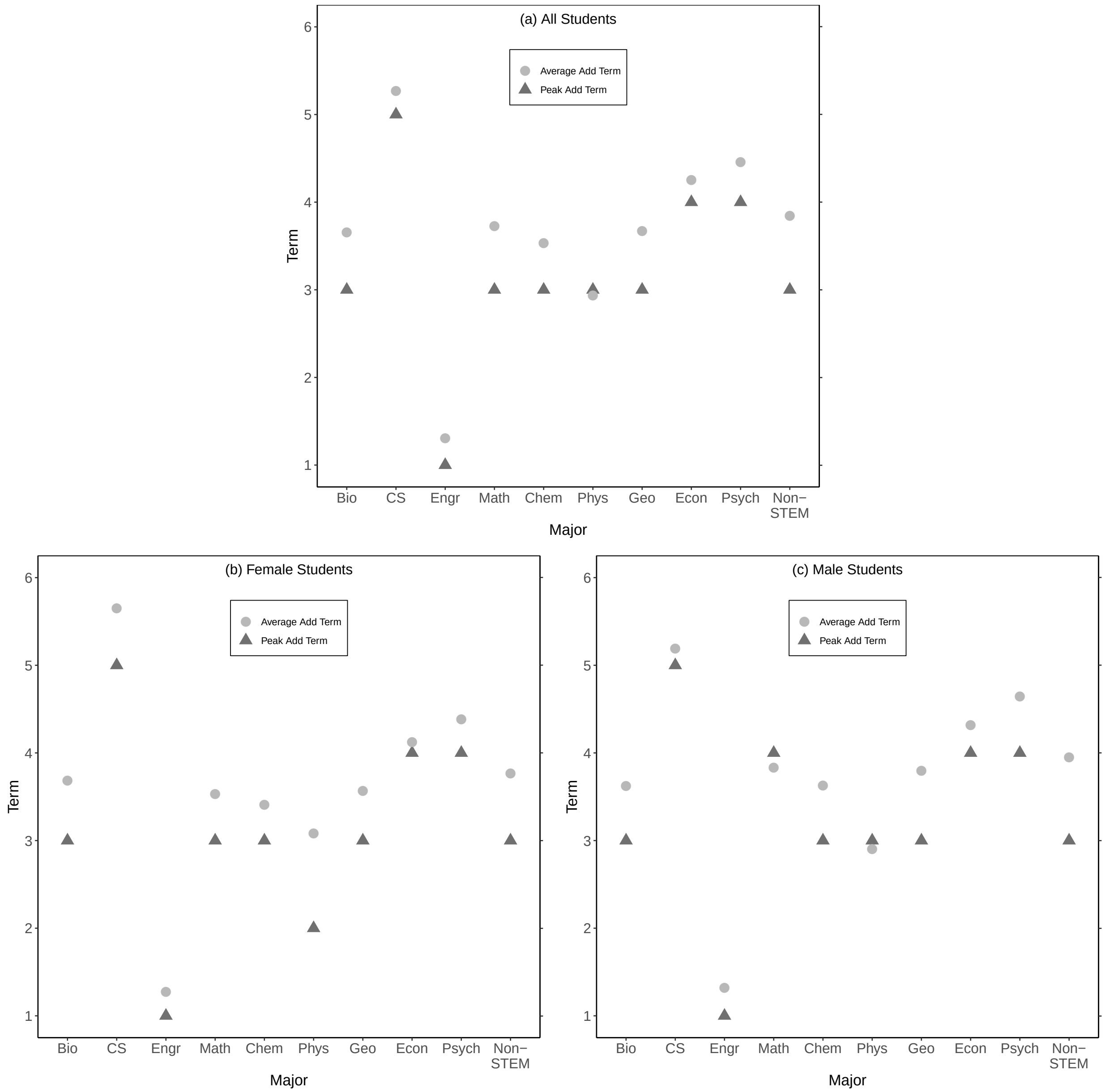


Figure 20: For each major, the term with the peak number of students adding the major in that term is plotted (triangles) as well as the average term in which students add that major (circles). This is done separately for (a) all students, (b) female students, and (c) male students.

For the majority of majors in Fig. 20a, the peak of students adding the major is in the third term (that is, the start of their second year), with an average between the third and fourth terms. The two non-STEM majors that we consider on their own, economics and psychology, depart slightly from this general trend, each with a peak in the fourth term and an average between the fourth and fifth terms. Two other majors, computer science and engineering, depart more significantly from the general trend in ways that can be explained by their particular implementation at the studied university.

Engineering has a peak in the first term in Fig. 20a, with an average only slightly after the first term. Since all students who enroll in the School of Engineering are considered "undeclared engineering" majors (the specific sub-discipline within engineering is not assigned in the first year), the majority of engineering students can be identified in their first term. Computer science instead has the latest peak term in Fig. 20a, namely in the fifth term, with an average slightly after the fifth term. This is due to the structure of the computer science program at the studied university, which does not allow students to declare the major until they have completed five of the required courses for the major. These trends in engineering and computer science are important to keep in mind while considering the results presented later in this paper, since in computer science we are not able to capture attrition that occurs (of students intending to major) during the terms before a student officially declares a major. Conversely in engineering, we are able to capture almost all attrition in the first year due to the unique enrollment conditions of engineering students, which is not possible for majors within the School of Arts and Sciences (where students can declare their major at any time after the first year).

Turning then to Figs. 20b and 20c, we see almost identical trends as in Fig. 20a. The two exceptions are that the peak declaration of physics majors for women occurs one semester earlier, in the second term (Fig. 20b). This is because the overall declaration of physics majors primarily occurs in two term, the second and third, which happens to result in different peaks but similar averages for men and women. Similarly, the peak declaration of mathematics majors for men occurs one semester later in the fourth term (Fig. 20c), again since mathematics majors overall are most likely to declare in the third or fourth term. Apart from these minor differences, these trends in major declaration term between men and women are virtually identical.

A more detailed accounting of the number of students that enroll in each term for each major is reported in Tables 23, 24, and 25 in Appendix C.1. Also, summaries of total number of unique students as well as the peak term and number of concurrent students in each major, students adding each major, and students dropping each major are available in Tables 26, 27, and 28 in Appendix C.2.

8.3.2 Attrition Rates

In order to answer **RQ2**, we further considered patterns of attrition rates by gender. In Fig. 21, we consider the drop rates of different subsets of students (all students, male students, and female students) in each major or group of majors. In Fig. 21a, we see that computer science, non-STEM, and psychology students are the least likely to drop their major, while physics, mathematics, and chemistry students are the most likely to drop. We note that the relatively low drop rate of computer science majors could be due to the late declaration of the computer science major seen in Fig. 20. That is, attrition from computer science prior to when students are allowed to declare the major is not accounted for in Fig. 21.

Fig. 21b shows the drop rates by gender, that is, separately for men and women. Though the drop patterns of men and women largely mimic the overall patterns in Fig. 21a, there are a few exceptions. In particular, in physics women are less likely to drop than men, with 31% of female physics majors dropping the major (Fig. 21b) and 40% of men dropping the major (Fig. 21c). Similarly, only 18% of female economics majors drop that major, in contrast with 27% of male economics majors drop that major.

8.3.3 Trajectories of Students After Dropping a Major

After discussing how many students drop each major, we answer **RQ3** by plotting in Fig. 22 where those dropped majors ended up. In particular, the major indicated in the legends of Fig. 22a and 22b shows which major was dropped, while the plot shows the percentage of those who dropped that major and ultimately earned a degree in each of the

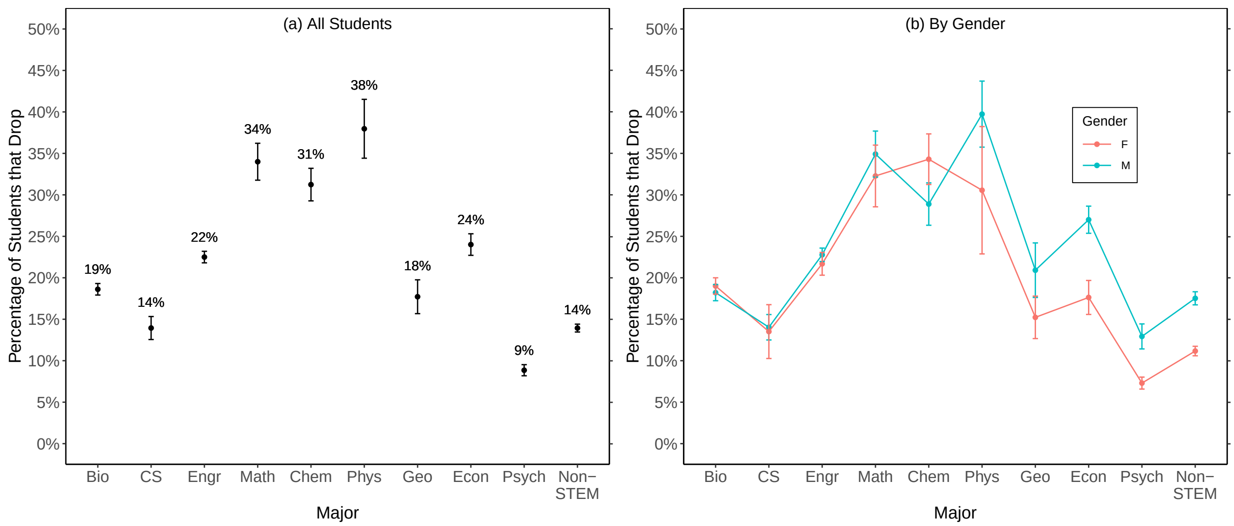


Figure 21: For each major, the percentage of students who declared the major but subsequently dropped the major is plotted along with its standard error. This is done separately for (a) all students and (b) men and women separately, along with lines connecting different points as guides to the eye.

majors on the horizontal axis, including the case when"no degree" was earned. For example in Fig. 22a, we see that among the students that drop the physics major (indicated by the line color in the legend), roughly 15% of them end up earning a degree in mathematics (by looking at this line's value above "Math" on the horizontal axis). The figure also shows that the two most common destinations for those who drop any major are either no degree or a degree in non-STEM, except for non-STEM majors who are most likely to earn no degree or a degree in psychology (another non-STEM major).

Apart from dropped STEM majors later earning degrees in non-STEM or leaving the university without a degree, we see a few other interesting spikes. For instance, those who drop a physics major are likely to earn a degree in mathematics (Fig. 22a) and those who drop chemistry or physics (Fig. 22a) as well as biological science (Fig. 22b) are likely to earn engineering degrees. Further, those who drop from economics are likely to major in mathematics (Fig. 22b). While all students who drop any major are very likely to earn no degree, the percentage of dropped majors in this category exceeds 50% for computer science (Fig. 22a), non-STEM, and psychology.

In order to further answer **RQ3**, Fig. 23 plots these same proportions of degrees earned by students who drop a major separately for men (Figs. 23a and 23b) and women (Figs. 23c and 23d). We see for the most part very similar patterns between men and women, with a few notable differences. For example, among students who drop a chemistry degree, we see that roughly 53% of the women eventually earn a degree in non-STEM (not including psychology or economics; Fig. 23c) compared with roughly 35% of the men (Fig. 23a). We see a similar pattern with the roles reversed among those students who drop a biology major, with roughly 15% of the men earning a degree in engineering (Fig. 23b) compared with less than 5% of the women (Fig. 23d). Another example is that men are more likely than women to earn computer science degrees after dropping a chemistry major (Figs. 23a and 23c). Finally, we note that across all of Fig. 23 in every major except psychology, the women who drop that major are either equally or more likely than the men to earn a degree in another major rather than leaving the university (that is, the women have a lower rate of earning "No Degree").

A more detailed accounting of the degrees earned by students who drop each major is

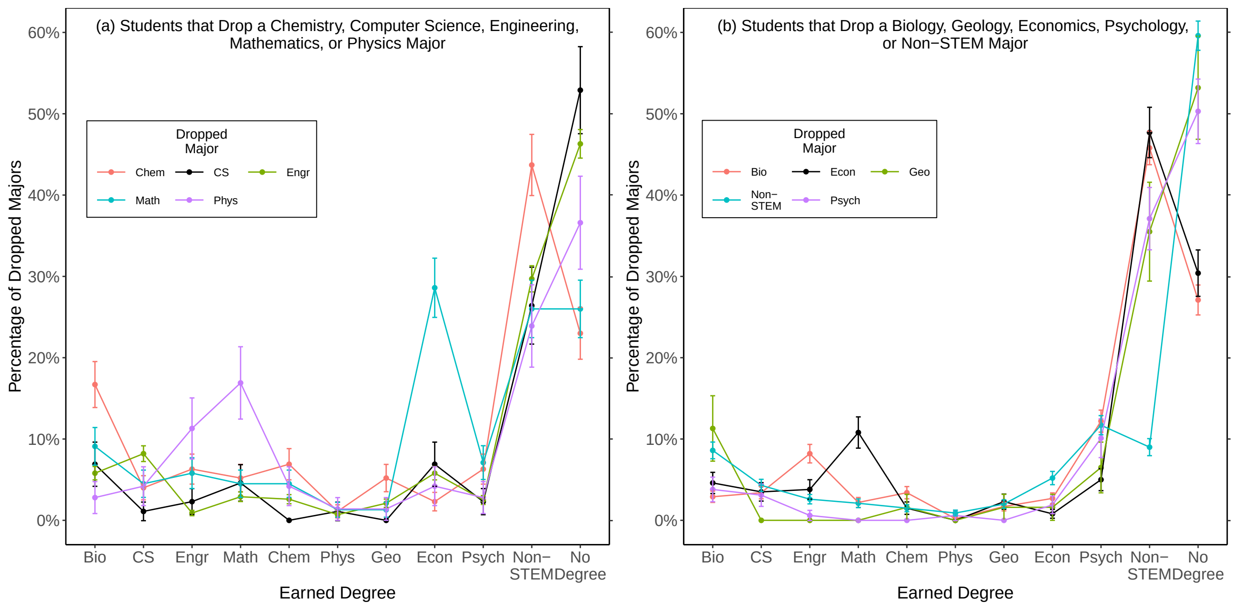


Figure 22: Among the students that drop each STEM major as well as psychology and non-STEM majors, the fractions of students that go on to earn a degree in other majors, or who do not earn a degree at all, are plotted along with their standard error. Dropped majors are grouped into (a) chemistry, computer science, engineering, mathematics, and physics and astronomy majors, and (b) biological science, geology, economics, psychology, and non-STEM majors.

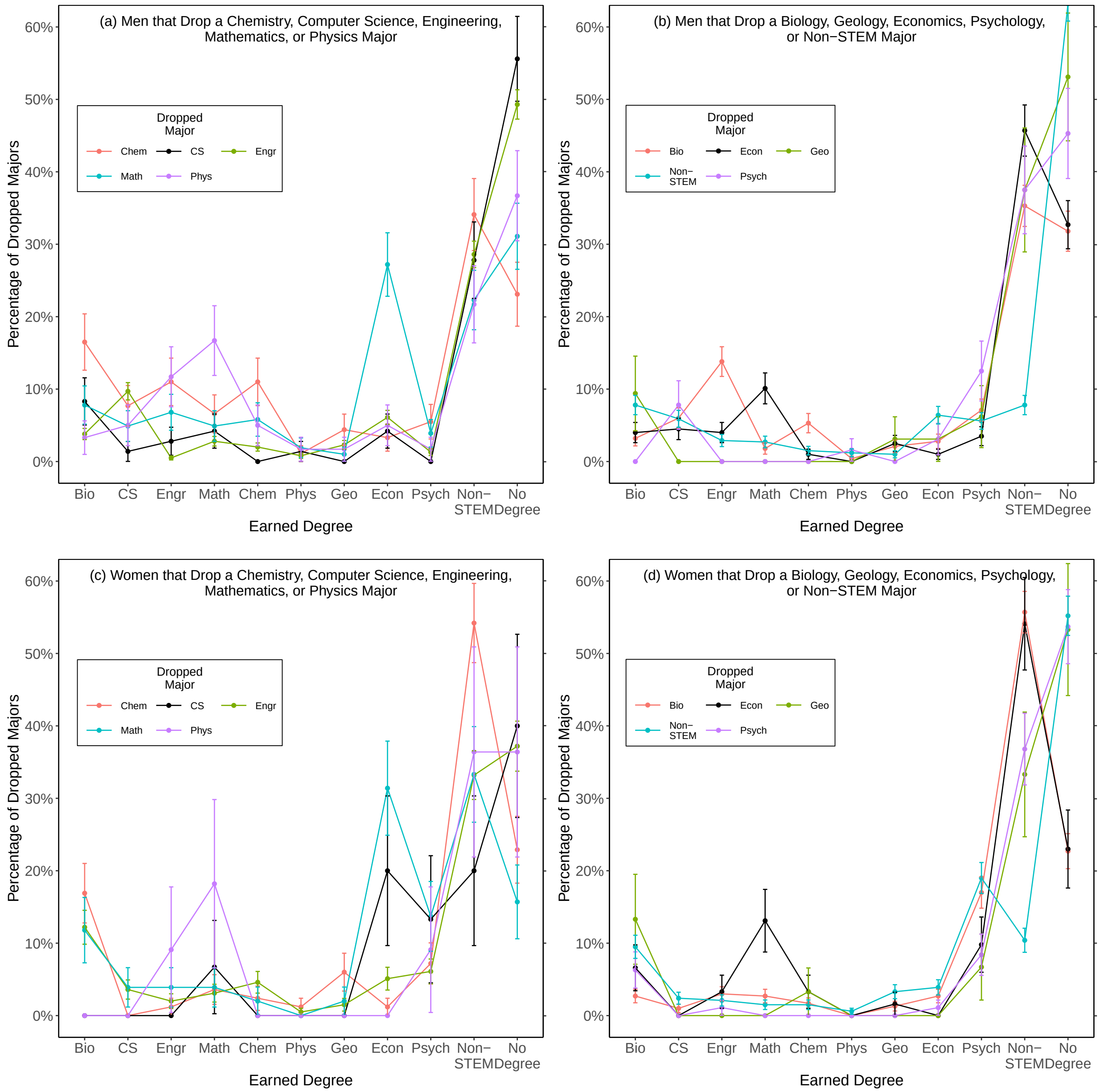


Figure 23: Among the men and women that drop each STEM major as well as psychology and other non-SEM majors, the percentages of men and women that go on to earn a degree in other majors, or who do not earn a degree at all, are plotted along with their standard error. Dropped majors are grouped into chemistry, computer science, engineering, mathematics, and physics and astronomy majors who are (a) men and (c) women, and biological science, geology, economics, and psychology majors who are (b) men and (d) women.

provided in Tables 29, 30, and 31 in Appendix C.3.

8.3.4 Degree-Earning Rates

In order to answer **RQ4**, we investigated how many students successfully earn a degree in each major. Figure 24a shows these degree-earning rates for all students in each major, while Fig. 24b shows these rates for female students and Fig. 24c for male students. While these are broadly similar to the reciprocal of the drop rates in Fig. 21, since some students drop a major and subsequently declare the same major again, these degree-earning rates are a more direct measurement of persistence in a major.

Looking first at the overall rates in Fig. 24a, there are fairly wide differences across majors, from the lowest rate in physics of about 65% to the highest in psychology and non-STEM, each at about 94%. The highest degree-earning rate in STEM occurs in computer science, with about 88% of declared computer science majors completing the degree requirements. As in Fig. 21, this can be at least partially explained by the requirements prior to declaring the major, which causes only students who have already progressed through a significant portion of the computer science curriculum to declare a computer science major.

Considering then the differences for women (Fig. 24b) and men (Fig. 24c), we see relatively few gender differences in these degree-earning rates. The slightly higher completion rate of women in non-STEM and psychology or men in chemistry, though statistically significant, are only differences of about 4-6%. As in Fig. 21, the largest difference between men and women seen here is in physics, with 75% of female physics majors earning a physics degree compared to 63% of male physics majors. However, the large error on these proportions, driven by the low sample size in physics shown in Fig. 18, makes it difficult to draw any conclusions from this gender difference in physics degree-earning rates. Similarly, women are more likely to complete a degree in economics, but again the size of the standard error prevents any conclusive statements about this difference.

Across all of Fig. 24, we note that since we have combined many majors for the "non-STEM" category, this is only a measure of the number of non-STEM majors who successfully earn a degree in any non-STEM major. That is, a student who drops one non-STEM major

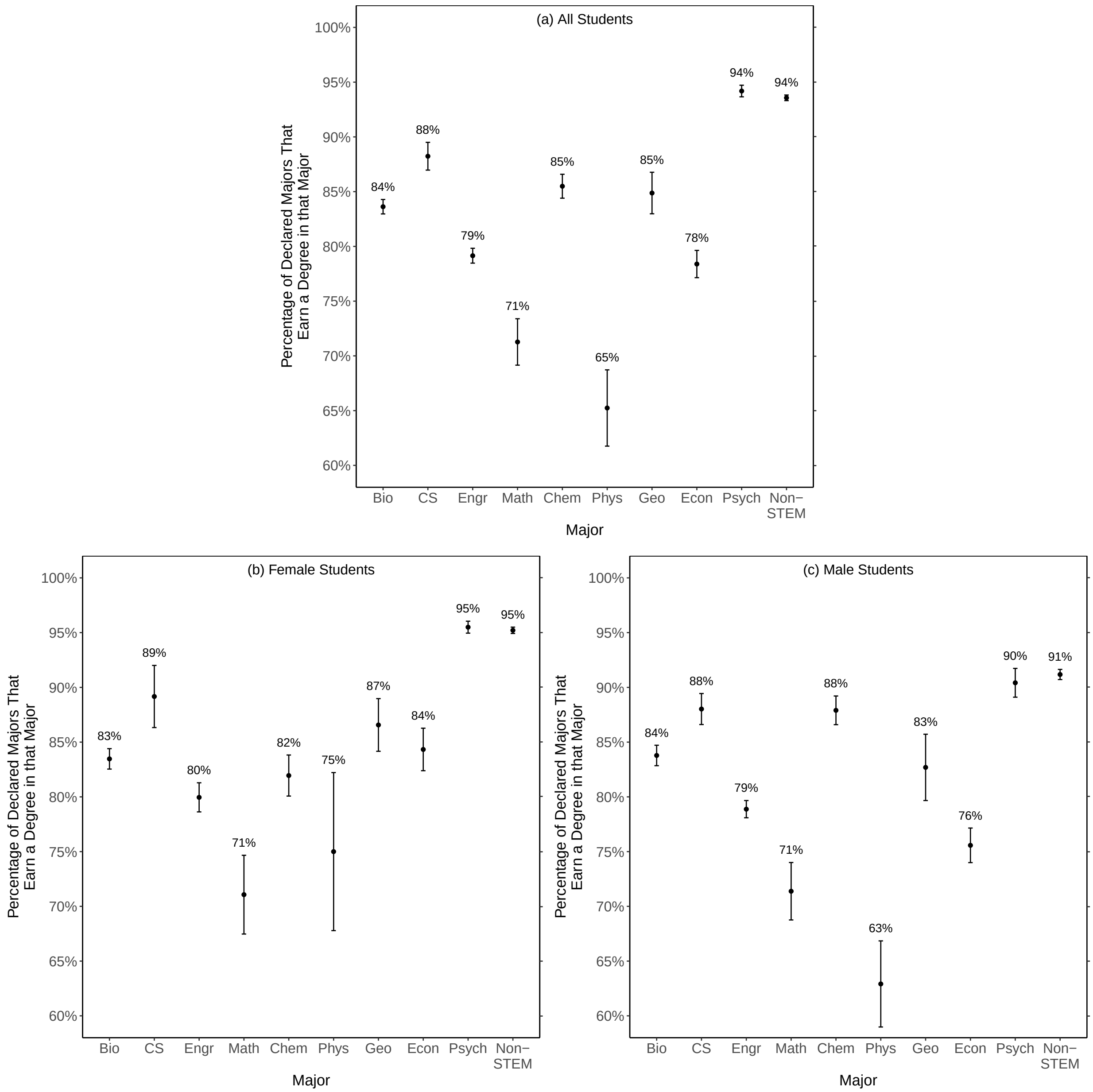


Figure 24: For each major listed on the horizontal axis, the percentages of (a) all students, (b) female students, and (c) male students who declare that major and then earn a degree in that major are plotted along with the standard error.

but earns a degree in a different non-STEM major will still be counted as having successfully earned a non-STEM degree. The same is true for the "engineering" category which also combine several majors. The high "success rates" of computer science and psychology may be due in part to the structure of their program encouraging students to declare slightly later than other disciplines, and so this measure may not be capturing attrition that happens prior to an official declaration of major (e.g., a student intending to major in a discipline decides against it before ever declaring that major). On the other hand, since all students enrolled in the engineering school are considered "undeclared engineering," the relatively low degree-earning rate of engineering reflects attrition even from the first to the second term, which is not captured for many other majors in which most students have not yet formally declared a major in their first term. Thus, each reported degree-earning rate here is a ceiling on the true rate that would include those students who intended to major but never declared.

8.3.5 Mean GPA of Degree-Earners vs. Major-Droppers

In order to further our understanding of why students may have dropped a given major and answer **RQ5**, Fig. 25 plots the mean GPA of STEM students who declared different sets of majors and then either earned a degree within that set of majors or dropped those majors. Note that students who dropped a major could have gone on to earn a degree with a different major or left the university without a degree. Both overall GPA (Figs. 25a, 25c, and 25e) and STEM GPA (Figs. 25b, 25d, and 25f) are plotted. Across all of Fig. 25, the large drop in sample size from year four to five and again from five to six is primarily due to students graduating. In Fig. 26 we similarly plot the overall GPA (Fig. 26a) and STEM GPA (Fig. 26b) of non-STEM majors.

We observe that in general, the students who dropped any given major have a lower GPA and STEM GPA than students who earned a degree in that major. However, the difference between the two groups varies based on which cluster of majors we consider. For biological science and neuroscience majors (Figs. 25a and 25b) as well as mathematics and physical science majors (Figs. 25e and 25f), those that earned a degree have a GPA of roughly 0.3 to 0.6 grade points higher than those that dropped, with similar differences in STEM GPA.

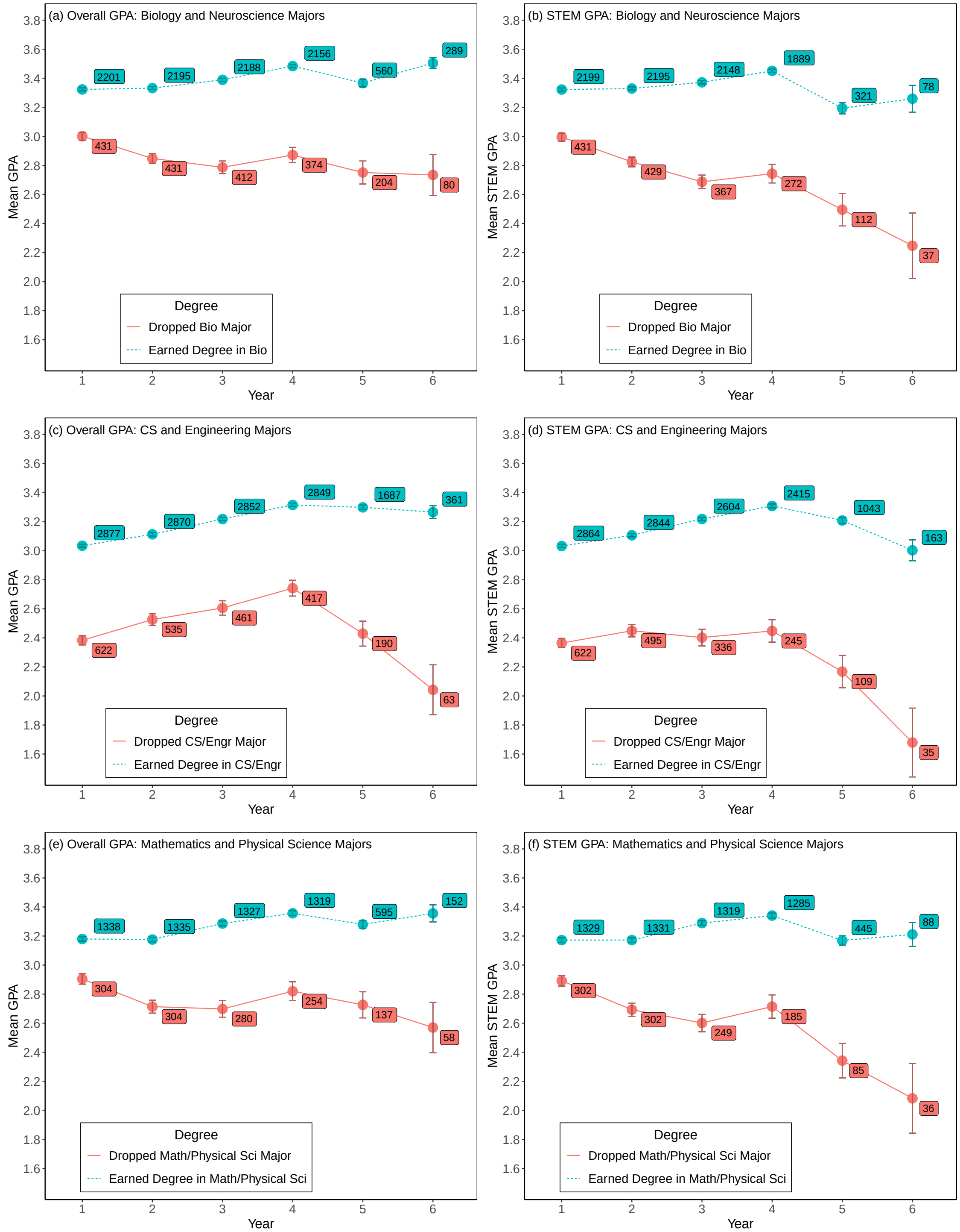


Figure 25: Mean GPA [(a), (c), and (e)] and STEM GPA [(b), (d), and (f)] along with standard error over time for STEM majors. Sample size listed by each point and guides to the eye connecting the points. Majors listed in plot legend entries.

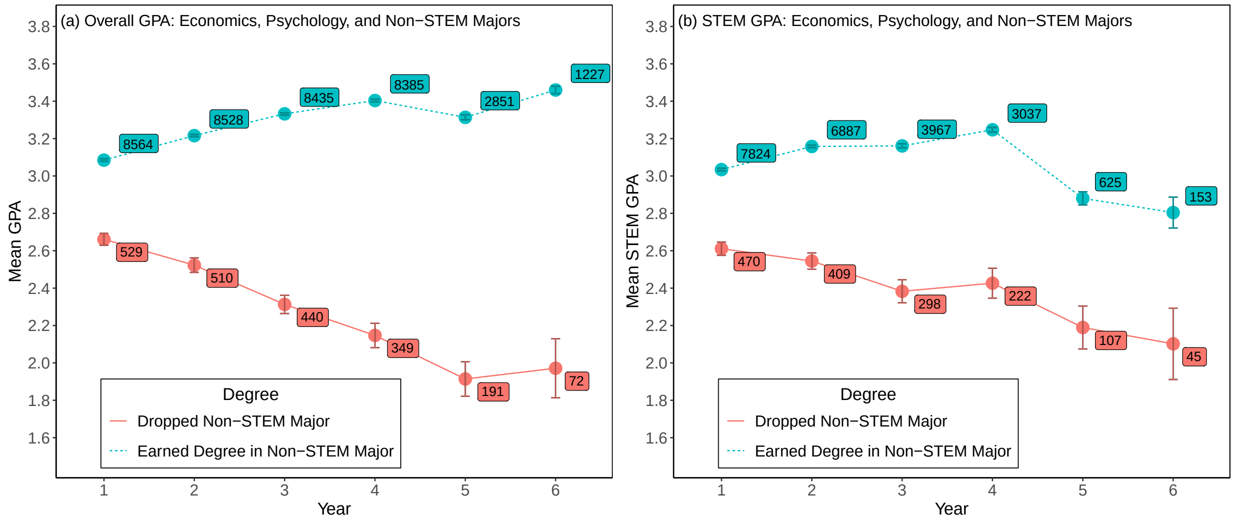


Figure 26: Mean GPA [(a), (c), and (e)] and STEM GPA [(b), (d), and (f)] along with standard error over time for non-STEM majors (including economics and psychology). Sample size listed by each point and guides to the eye connecting the points.

This difference in grade points represents a difference of one to two letter grades at the studied university, where, for example, the difference between a B and B+ is 0.25 grade points and between a B+ and A− is 0.5 grade points. For computer science and engineering majors (Figs. 25c and 25d), those that earned a degree in this set of majors have a GPA of roughly 0.6 grade points higher than those that dropped. Finally, for non-STEM majors including psychology (Fig. 26), the overall GPA disparity widens over time from roughly 0.4 grade points in the first year to roughly 1.2 grade points in the fourth year, while in STEM courses the GPA disparity rises from roughly 0.4 grade points in the first year to roughly 0.8 grade points in the fourth year. Notably, a much smaller fraction of students are dropping from this set – about 6% of the total number of students – which could be due in part to the wide net of considering all non-STEM majors together.

We further consider the same measures separately for men and women in STEM in Fig. 27, with the same subfigure structure as Fig. 25. Similarly, we consider the GPA of men and women in non-STEM in Fig. 28, with the same subfigure structure as Fig. 26. In all cases,

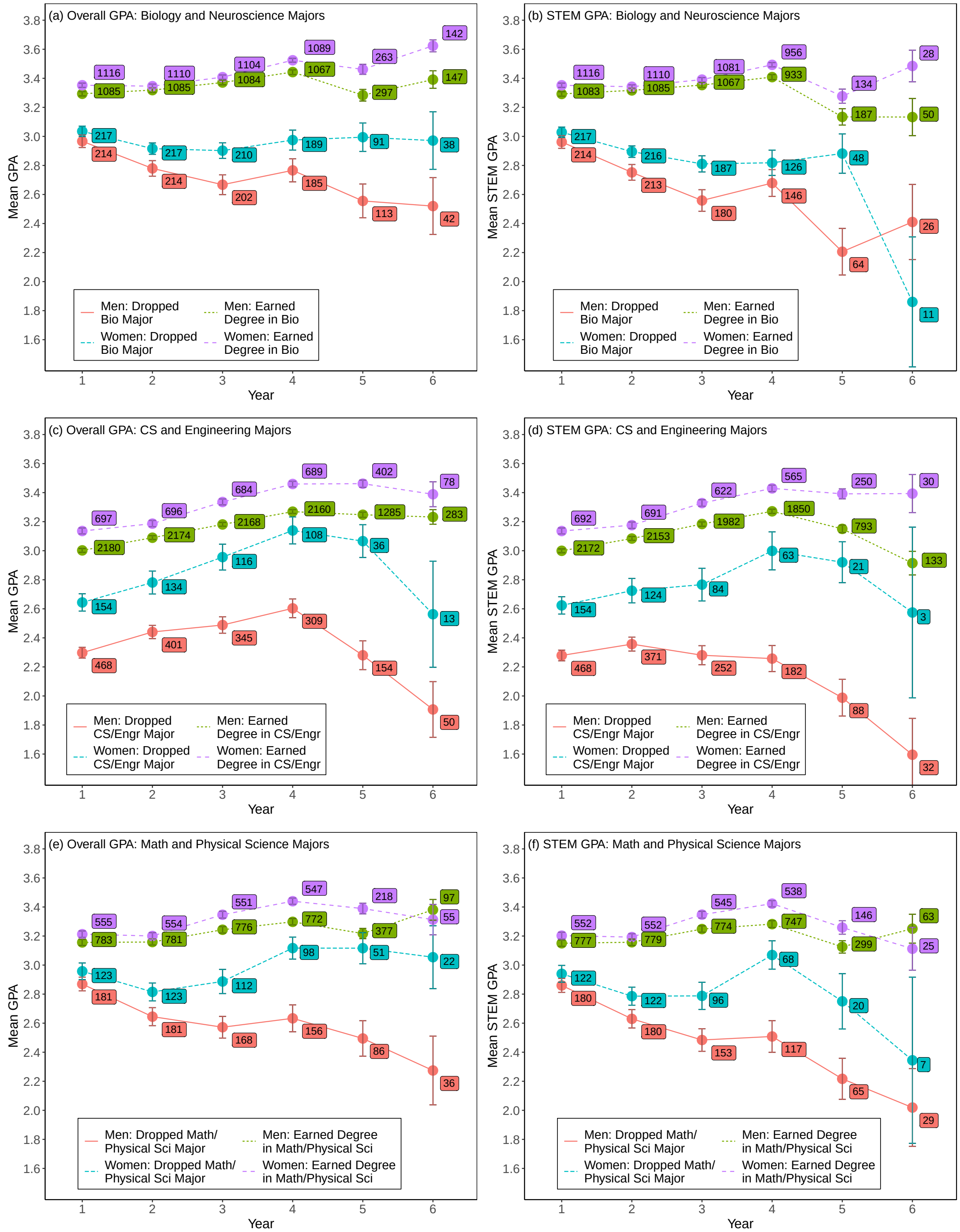


Figure 27: Mean GPA [(a), (c), and (e)] and STEM GPA [(b), (d), and (f)] along with standard error over time for STEM majors by gender. Sample size listed by each point and guides to the eye connecting the points. Majors listed in plot legend entries.

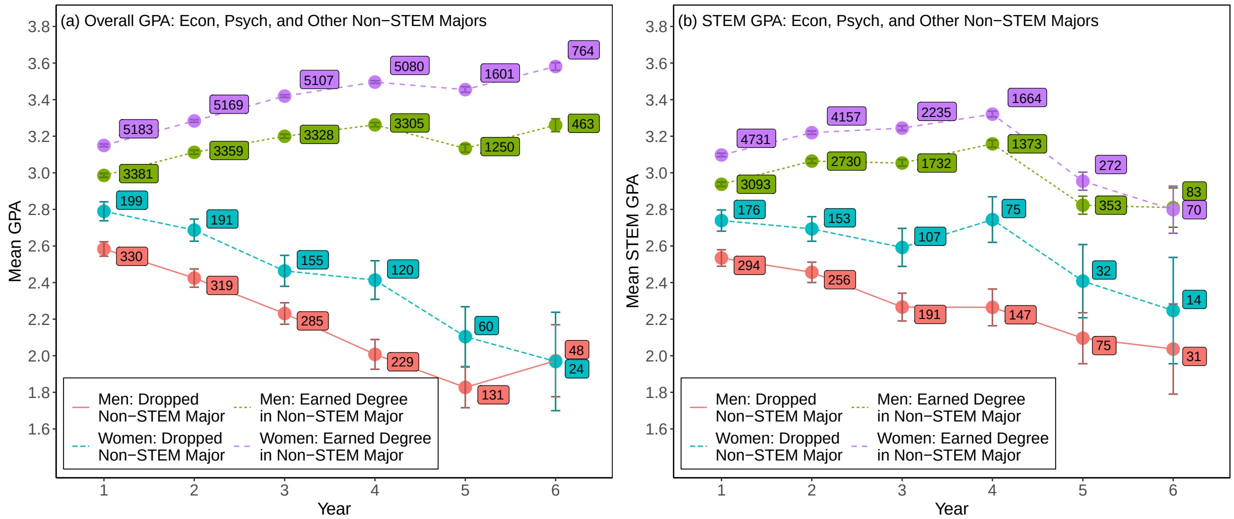


Figure 28: Mean GPA [(a), (c), and (e)] and STEM GPA [(b), (d), and (f)] along with standard error over time for non-STEM majors (including economics and psychology) by gender. Sample size listed by each point and guides to the eye connecting the points.

the main finding is that women are earning higher grades on average than men, both among those who drop a given major and those who earn a degree in that major. However, the grade differences between men and women who dropped the major and men and women who earned a degree in the major are inconsistent. Across all majors in Fig. 27, the differences between men and women who drop a given major are larger than the differences between men and women who earn a degree in that major. This is particularly noticeable in Figs. 27c, 27d, 27f, and 27e, where by the fourth year the women who have dropped these majors (computer science, engineering, mathematics, and physical science) are earning nearly the same GPA and STEM GPA as the men who persisted in these majors.

In order to test these differences, we report in Tables 18a and 18b the results from four comparisons. We compare the STEM GPA of the women who drop each set of STEM majors (Figs. 27b, 27d, and 27f) with the STEM GPA of men who drop those same majors (Table 18a) and women who earn a degree in the STEM majors with men who earn a degree in the same majors (Table 18b). These comparisons are performed with two measures: the

p-value from a two-tailed t-test[98] and the effect size from Cohen's d [65]. These tests were not performed for non-STEM majors since such a low percentage of the students, roughly 6%, dropped from the non-STEM majors (including psychology) altogether (Fig. 26).

In each set of majors, we see that the mean STEM GPA difference between women and men who drop each set of STEM majors has a small to medium effect size (d between 0.30 and 0.56), while the STEM GPA difference between women and men who persist in the majors has only a small effect size (d between 0.10 and 0.27). In both cases, the women are earning higher grades than the men. Thus, the women who are dropping these majors have, on average, better grades than their male counterparts by one to two letter grades while the women who persist earn very similar grades to the men who persist. This is indicative that women are more likely to drop the major than men with the same grades when those grades fall below the mean.

8.4 Discussion

In this section, we will begin by discussing the general trends (i.e., setting aside the gender differences), and then follow up with a discussion of the gender differences.

8.4.1 General Enrollment Patterns

Despite large differences in the number of students enrolling in different STEM disciplines at the studied university (Fig. 18), there are broadly similar patterns of when those students declare the major (Fig. 20a), with some exceptions (i.e., engineering and computer science due to the constraints on when a student can declare a major). However, there are notable differences in the attrition of students from the different majors (Fig. 21), and the corresponding degree-earning rates (Fig. 24a). Notably, a few STEM disciplines stand out as having particularly high rates of attrition (or low rates of degree completion for students who declared those majors), e.g., mathematics and physics. This is consistent with the Leslie *et al.* study which identifies mathematics and physics as the two STEM disciplines with the

Table 18: A comparison of the cumulative STEM GPA of women who drop STEM majors with (a) the men who drop the same STEM majors and (b) the men who persist and earn a degree in the same STEM majors. Along with the number of students (N), we report the mean (M) and standard deviation (SD) of cumulative STEM GPA for each group. The p-value from a two-tailed t-test is reported comparing the women and men, along with Cohen's d measuring the effect size of the gender difference (the sign of d matches the sign of the mean GPA for women minus the mean GPA for men). The comparison is performed separately for the STEM majors corresponding to the indicated figure, i.e., for Fig. 27b we consider biological science and neuroscience majors; for Fig. 27d we consider computer science and engineering majors; and for Fig. 27f we consider mathematics and physical science majors.

(a)	Women who Dropped the Major			Men who Dropped the Major			Statistical Comparisons	
Fig.	N	M	SD	N	M	SD	p	d
27b	217	2.95	0.52	214	2.76	0.73	0.002	0.30
27d	154	2.68	0.78	468	2.23	0.81	< 0.001	0.56
27f	176	2.91	0.64	266	2.66	0.74	< 0.001	0.36

(b)	Women who Earned the Degree			Men who Earned the Degree			Statistical Comparisons	
Fig.	N	M	SD	N	M	SD	p	d
27b	1116	3.40	0.40	1085	3.36	0.44	0.017	0.10
27d	696	3.26	0.46	2179	3.13	0.52	< 0.001	0.27
27f	909	3.30	0.44	1264	3.23	0.50	< 0.001	0.15

highest "ability belief" (i.e., emphasis on brilliance) [158]. This trend of high attrition rates is particularly problematic for mathematics and physics, since these disciplines recruit very few students in the first place (Fig. 18). Moreover, mathematics and physics are also two disciplines with deeply hierarchical knowledge structures, which could influence student decision making, e.g., whether to leave the discipline after unsatisfactory experiences in earlier courses.

8.4.2 Gendered Enrollment Patterns

The most notable example of gender differences in enrollment patterns observed in our analysis is in Fig. 19. In biological science, geological and environmental science, neuroscience, and statistics we see a balanced representation of men and women. However, we see an underrepresentation of women in chemistry, computer science, engineering, mathematics, physics, and economics, and a corresponding overrepresentation of women in non-STEM including psychology (Fig. 19a). Again, the results observed here are roughly consistent with those observed by Leslie *et al.* [158]. The gender imbalance in these STEM disciplines itself plays a pernicious role in recruitment and retention of women who do not have many role models and also affects the performance of women, who are constantly forced to prove themselves and counter the societal stereotypes working against them in these fields.

Furthermore, despite these differences in the number of men and women in these STEM disciplines, we see very few differences in the remainder of the enrollment measures, including the time of major-declaration (Figs. 20b and 20c), rates of attrition (Fig. 21b), and rates of degree-attainment (Figs. 24b and 24c). There are some hints towards differing rates of attrition in physics and economics but these differences suffer from large standard error, particularly in physics, due to a low sample size. Moreover, we hypothesize that the higher attrition rate of men in physics and economics may be due to the pressure that women experience not to enroll in these disciplines in the first place, which would have occurred prior to the declaration of majors and instead manifests in the stark underrepresentation of women in these disciplines noted earlier (Fig. 19a). There have been many studies that find highly problematic gender inequities in introductory physics and mathematics that could at least partly contribute to this underrepresentation [137, 177, 178, 181].

8.4.3 Trajectories of Students Who Drop a Major

As with the attrition and degree-earning rates, we see broadly similar patterns between men and women who drop the various STEM majors (Fig. 23). Moreover, we find that across all disciplines (including non-STEM majors) except psychology, women who drop a major are either equally or more likely to subsequently earn a degree in a different major at the same university than men who drop a major (see the "No Degree" entries in Fig. 23), who are more likely to leave the university altogether, either by dropping out of college completely or transferring to another university. This may be easier to see in Tables 30 and 31 in Appendix C.3.

8.4.4 Gendered GPA Differences

We find pervasive and deeply troubling gendered trends in the overall GPA and STEM GPA of those students who drop different STEM majors. In particular, Fig. 27 shows that the women who drop a STEM major have a higher average GPA than the men who drop a STEM major, and these differences are shown to be statistically significant as shown in Table 18a. This is in contrast with comparisons of men and women who persist and earn degrees in STEM majors (Table 18b), where women still have higher grades but with only a small effect size. Thus, on average, among students with the same GPA, the women in all STEM majors are more likely to drop the majors than the men, who are more likely to persist. It is important to note that this is true both among the STEM majors with gender imbalances in the population as well as those majors without imbalances.

While it is true that women at the studied university on average earn higher grades than men in most STEM courses, that does not explain why women are choosing to drop STEM majors with the same grades as men who persist. The difference must then be coming from another source, for example inequitable and non-inclusive learning environments and lack of mentoring and support for women who may have lower self-efficacy [8, 9, 10, 11, 12, 13]. Women may also have a lower sense of belonging and value pertaining to remaining in these disciplines [88, 87, 86, 236] if learning environments are not equitable and inclusive, especially because they must bear the high cost of managing the burden of societal stereotypes and the ensuing stereotype threats. We hypothesize that the brilliance attributions of STEM disciplines [158] and who is likely to excel in them could influence women away from a discipline in which they could have succeeded. Thus, without explicit effort to improve the learning environments in these disciplines, the culture and stereotypes surrounding STEM in general may be creating an environment in which women are being unfairly driven out of these fields in which they could have thrived while their male counterparts are not subjected to these same pressures and are persisting with worse performance.

8.4.5 Limitations and Implications

One limitation of this study is that physics, which has the most consistently problematic gender differences, also has the lowest number of students. Future studies can make use of more data (either data available further back in time or as more data continue to accumulate) to explore the gender differences in physics better. This study also limits its considerations to gender. Other studies with larger data sets could investigate how other underserved populations are being left behind in STEM, such as underrepresented minority students, first-generation college students, or low-income students.

A critically important extension of this work would be for other institutions of different types and sizes to do similar analyses in order to broaden the wealth of knowledge available and continue to work towards the goal of equitable and inclusive education. Other institutions noting similar highly problematic trends can help pinpoint common sources of inequities, while institutions that do not observe these trends may be able to identify how they have structured their programs to avoid these inequitable trends. Studies such as this can thus provide a framework for other institutions to perform similar analyses, and for particular departments to understand how their own trends differ from those of other departments at their own university. For instance, our findings here for, e.g., physics and mathematics drop out rates indicate that there is substantial room to improve their support of their intended majors so they do not drop out. Focus on increasing equity and inclusion in learning is especially important in the early courses for these majors, since they are fraught with problematic gender differences and may be contributing to the underrepresentation of women in these majors in the first place. Similar steps should be taken for the other majors that have low representation of women, especially computer science and engineering, and to a lesser extent chemistry.

The other STEM majors (biological sciences, economics, geology and environmental science, neuroscience, and statistics) should also take a closer look at which students are choosing to leave those disciplines, and why women choosing to leave have similar GPAs as the men who persist and earn degrees in those disciplines. These are signs of inequity even in majors in which women are not underrepresented. All of these issues should be addressed since they are critical for improving equity and inclusion in higher education STEM learning environments.

8.5 Acknowledgments

This research is supported by the National Science Foundation Grant DUE-1524575 and the Sloan Foundation Grant G-2018-11183.

9.0 Inequity in Grades and Persistence: Race/Ethnicity

9.1 Introduction and Theoretical Framework

Increasingly, Science, Technology, Engineering, and Mathematics (STEM) departments across the US are focusing on using evidence to improve the learning of all students, regardless of their background and making learning environments equitable and inclusive [219, 136, 137, 182, 188, 194, 151, 172, 171, 186, 108, 39, 37, 38, 135, 119, 118, 72, 40]. However, racial/ethnic minority students are still severely underrepresented in many STEM disciplines [199]. In order to understand the successes and shortcomings of the current state of education, the use of institutional data to investigate past and current trends is crucial. In the past few decades, institutions have been keeping increasingly large digital databases of student records. We have now reached the point where there are sufficient data available at many institutions for robust statistical analyses using data analytics that can provide invaluable information for transforming learning environments and making them more equitable and inclusive for all students [6, 217]. Studies utilizing many years of institutional data can lead to analyses that were previously limited by statistical power. This is particularly true for studies of performance and persistence in STEM programs that rely on large sample sizes [212, 166, 92, 172, 192, 167, 211, 181, 283, 151, 236, 171, 186].

In this study, we use 10 years of institutional data from a large state-related research university to investigate how patterns of student major-declaration and subsequent degree-earning may differ by self-reported race/ethnicity. We will consider the two racial/ethnic groups that are over-represented in US Bachelor's degree attainment [199], namely Asian and White students, separately from other students whom we call underrepresented minority (URM) students. The theoretical framework for this study has two main foundations: critical theory and expectancy value theory.

Critical theories, e.g., of race and gender, focus on historical sources of inequities within society, i.e., societal norms that perpetuate obstacles to the success of certain groups of disadvantaged people [106, 69, 150, 286, 110, 258, 261, 242, 188]. Critical theory tells us that the dominant group in a society perpetuates these norms, which are born out of their interests, and pushes back against support systems that seek to subvert these norms [69, 150, 286]. In our case, critical race theory provides a historical perspective on the much-studied racial/ethnic inequities in STEM. Much important work has been done that relates to critical theory of race and ethnicity in STEM education [3, 284, 7, 137, 252, 160, 14, 213, 261, 107, 198, 245, 242], e.g., pervasive stereotypes about who can and cannot succeed in STEM puts additional burden on URM students in STEM courses.

Expectancy value theory (EVT) is another framework that is central to our investigation and states that a student's persistence and engagement in a discipline are related to their expectancy about their success as well as how the student values the task [88, 87, 86]. In an academic context, "expectancy," which refers to the individual's beliefs about their success in the discipline, is closely related to Bandura's construct of self-efficacy, defined as one's belief in one's capability to succeed at a particular task or subject [8, 9, 10, 11, 12, 13, 88, 87, 86].

There are four main factors that influence students' expectancy or self-efficacy, namely vicarious experiences (e.g., instructors or peers as role models), social persuasion (e.g., explicit mentoring, guidance, and support), level of anxiety, and performance feedback (e.g., via grades on assessment tasks) [8, 9, 10, 11, 12, 13]. URM students generally have lower self-efficacy than non-URM students in many STEM disciplines because these four factors negatively influence them [4, 70, 31, 46, 29, 15, 126]. For example, URM students are underrepresented in their classrooms across STEM, and less likely to have a role model they can relate to among the faculty [199]. Further, the stereotypes surrounding URM students in many STEM disciplines can affect how they are treated by mentors, even if such an effect is subconscious [4, 70, 31, 46, 29, 15, 126]. Moreover, URM students are susceptible to stress and anxiety from stereotype threat (i.e., the fear of confirming stereotypes about URM students in many STEM disciplines) which is not experienced by their non-URM peers [4, 70, 31, 46, 29, 15, 126]. This stress and anxiety can rob them of their cognitive resources, especially during high-stakes assessments such as exams.

Expectancy can influence grades earned as well as the interest and likelihood to persist in a program [8, 9, 10, 11, 12, 13, 272]. Stereotype threats that URM students in many STEM disciplines experience can increase anxiety in learning and test-taking situations and lead to deteriorated performance. Since anxiety can increase when performance deteriorates, these factors working against URM students in STEM can force them into a feedback loop and hinder their performance further, which can further lower their self-efficacy and can continue to affect future performance [8, 9, 10, 11, 12, 13, 272].

In EVT, value is typically defined as having four facets: intrinsic value (i.e., interest in the task), attainment value (i.e., the importance of the task for the student's identity), utility value (i.e., the value of the task for future goals such as career), and cost (i.e., opportunity cost or psychological effects such as stress and anxiety) [88, 87, 86]. In the context of URM students' enrollment and persistence in many STEM disciplines, the societal stereotypes can influence all facets of the students' value of these STEM disciplines. Intrinsic value can be informed by societal stereotypes surrounding the STEM disciplines, and attainment and utility values can be further tempered by these stereotypes. Utility value is an important facet of student education in STEM, since a degree in a STEM field provides many job opportunities for graduating students. In addition, the psychological cost of majoring in these disciplines can be inflated by the stereotype threat. All of these effects can conspire to suppress the likelihood of students choosing and/or persisting in various STEM disciplines [272].

In order to measure the long-term effects of these systemic disadvantages, we investigate the differences in attrition rates and choices of major of Asian, URM, and White students over the course of their studies at one large public research university using 10 years of institutional data. Since these disadvantages to students can be context-dependent, we will consider the attrition rates in many different STEM majors and non-STEM majors in order to understand the trends in each discipline.

9.1.1 Research Questions

Our research questions regarding the relationships between race/ethnicity and persistence in a degree are as follows.

RQ1. How many students major in each discipline? How many Asian, URM, and White students major in each discipline?

RQ2. Do rates of attrition from the various majors differ? Do rates of attrition from the various majors differ for Asian, URM, and White students?

RQ3. Among those students who drop a given major, what degree, if any, do those students earn? How do these trends differ for Asian, URM, and White students?

RQ4. What fraction of declared majors ultimately earn a degree in that major in each STEM subject area? How do these trends differ for Asian, URM, and White students?

RQ5. What are the GPA trends over time among students who earn a degree in a given major and those who drop that major? How do these trends differ for Asian, URM, and White students?

9.2 Methodology

9.2.1 Sample

Using the Carnegie classification system, the university at which this study was conducted is a public, high-research doctoral university, with balanced arts and sciences and professional schools, and a large, primarily residential undergraduate population that is full-time and reasonably selective with low transfer-in from other institutions [133].

The university provided for analysis the de-identified institutional data records of students with Institutional Review Board approval. In this study, we examined these records for $N = 18,009$ undergraduate students enrolled in two schools within the university: the School of Engineering and the School of Arts and Sciences. This sample of students includes all of those from ten cohorts who met several selection criteria, namely that the student had first enrolled at the university in a Fall semester from Fall 2005 to Fall 2014, inclusive, and the student had either graduated and earned a degree, or had not attended the university for at least a year as of Spring 2019. This sample of students is 49.8% female and had the following race/ethnicities: 77.8% White, 11.1% Asian, 6.9% Black, 2.5% Hispanic, and 1.8% other or multiracial.

9.2.2 Measures

9.2.2.1 Race/Ethnicity The institutional data provided by the university included self-reported race/ethnicity of the students. Students were asked to indicate all of the following racial/ethnic groups with which they identified: American Indian/Alaskan, Asian, Black, Hawaiian/Pacific Islander, Hispanic, White, or Other. The researchers converted the student answers to this question into three categories for analysis, defined as follows.

- "Asian": Students who selected only Asian or only Asian and White.
- "White": Students who selected only White.
- "URM": Underrepresented minority students who selected any option(s) other than Asian or White (including those who additionally selected Asian or White).

A very small number of students chose not to answer this question, and were removed from analysis.

9.2.2.2 Academic Performance Measures of student academic performance were also included in the provided data. High school GPA was provided by the university on a weighted scale from 0-5 that includes adjustments to the standard 0-4 scale for Advanced Placement and International Baccalaureate courses. The data also include the grade points earned by students in each course taken at the university. Grade points are on a 0-4 scale with A = 4, B = 3, C = 2, D = 1, F = 0, where the suffixes "+" and "−" add or subtract, respectively, 0.25 grade points (e.g. B− = 2.75), with the exception of A+ which is reported as the maximum 4 grade points. The courses were categorized as either STEM or non-STEM courses, with STEM courses being those courses taken from any of the following departments: biological sciences, chemistry, computer science, economics, any engineering department, geology and environmental science, mathematics, neuroscience, physics and astronomy, and statistics. We note that for the purposes of this paper, "STEM" does not include the social sciences other than economics, which has been included due to its mathematics-intensive content.

9.2.2.3 Declared Major and Degree Earned For each student, the data include their declared major(s) in each semester as well as the major(s) in which they earned a degree, if any. The data were transformed into a set of binary flags for each semester, one flag for each possible STEM major as well as psychology and a general non-STEM category for all other majors. A similar set of flags was created for the degrees earned by students. From these flags, we tabulated a number of major-specific measures in each semester, including

- current number of declared majors,
- number of newly declared majors from the preceding semester, and
- number of dropped majors from the preceding semester.

The total number of unique students that ever declared or dropped a major were also computed. The subset of students that dropped each major was further investigated and the major in which they ultimately earned a degree, if any, was determined.

Throughout this paper we group the STEM majors into three clusters: biological sciences (including neuroscience) and neuroscience; computer science and engineering; and mathematics (including statistics), chemistry, physics and astronomy, and geology and environmental science. When ordering majors (i.e., in figures and tables), the majors will be presented in the order they are listed in the preceding sentence, followed by the non-STEM majors: economics, psychology, and "other non-STEM" (i.e., all other non-STEM majors). Note that "engineering" groups together all engineering majors for departments in the School of Engineering at the studied university. These majors include chemical, computer, civil, electrical, environmental, industrial, and mechanical engineering as well as bioengineering and materials science.

Finally, we will make use of shortened labels for the majors in figures and tables. These shortened labels are defined in Table 19.

9.2.2.4 Year of Study Finally, the year in which the students took each course was calculated from the students' starting term and the term in which the course was taken. Since the sample only includes students who started in fall semesters, each "year" contains courses taken in the fall and subsequent spring semesters, with courses taken over the summer

Table 19: A list of the majors considered in this study and the shortened labels used to refer to those majors in tables and figures. Note that "Engineering" is a combination of many engineering majors offered by the School of Engineering.

Major	Short Label
Biological Sciences and Neuroscience	Bio
Computer Science	CS
Engineering	Engr
Mathematics and Statistics	Math
Chemistry	Chem
Physics and Astronomy	Phys
Geology and Environmental Science	Geo
Economics	Econ
Psychology	Psych
Other Non-STEM	Non-STEM

omitted from this analysis. For example, if a student first enrolled in Fall 2007, then their "first year" occurred during Fall 2007 and Spring 2008, their "second year" during Fall 2008 and Spring 2009, and so on in that fashion. If a student is missing both a fall and spring semester during a given year but subsequently returns to the university, the numbering of those post-hiatus years is reduced accordingly. If instead a student is only missing one semester during a given year, no corrections are made to the year numbering.

9.2.3 Analysis

For each student, we calculate their grade point average (GPA) across courses taken in each year of study from their first to sixth years. In addition, we calculate the student's STEM GPA in each year, that is, the GPA in STEM courses alone. The mean GPA as well as the standard error of the mean is computed for various groupings of students [98].

Further, proportions of students in various groups (i.e., grouped by major and/or demographic group) are calculated along with the standard error of a proportion [98]. In particular, the proportions we report are

- the proportion of students in each major that are Asian, URM, or White students,
- the proportion of Asian, URM, and White students, respectively, that declare each subject as a major,
- the proportion of declared majors that drop the major,
- the proportion of those who drop each major that earn a degree in each other major, and
- the proportion of all declared majors that ultimately earn a degree in that major.

All analyses were conducted using R [226], making use of the package `tidyverse` [279] for data manipulation and plotting.

9.3 Results

9.3.1 Major Declaration Patterns

There are many angles with which we can approach **RQ1** and investigate patterns of student major declaration. First, Fig. 29 shows the number of students that ever declared each major. This is done both overall (Fig. 29a) and for Asian students (29b), URM students (29c), and White students (29d) separately. These results provide an important context for the subsequent analyses that may be partially explained by the number of students in each major.

Figures 29b and 29c begin to hint at differences in enrollment patterns for the different racial/ethnic groups. URM and White students (Figs. 29c and 29d) have relatively similar patterns, with some small differences such as a slightly higher fraction of White students declaring engineering majors and a slightly higher fraction of URM students declaring psychology majors. Asian students show some trends that differ from both URM and White students, in particular a higher fraction of Asian students major in biological sciences and economics. These patterns are explored further in Fig. 30 by standardizing the scales in two ways. In Fig. 30a, we consider the populations of each major individually and calculate the percentages of that population that consist of Asian, URM, or White students. This provides insight into what these students might be seeing in the classes for their major (i.e., the demographic makeup of the classrooms). The trends seen in Fig. 30a are largely dominated by the relative representation of Asian, URM, and White students in the university as a whole.

Another way to represent the population of these majors is to consider what percentage of all Asian, URM, or White students choose each major, as seen in Fig. 30b. While this plot mimics those in Figs. 29b, 29c, and 29d, we can now see the differences noted earlier more clearly. For example, in Fig. 30b we can again see that the majors chosen by URM and White students are similar, with the exceptions of engineering (chosen by slightly more White students) and psychology and other non-STEM (chosen by slightly more URM students, and not including economics which has similar enrollment percentages for URM and White

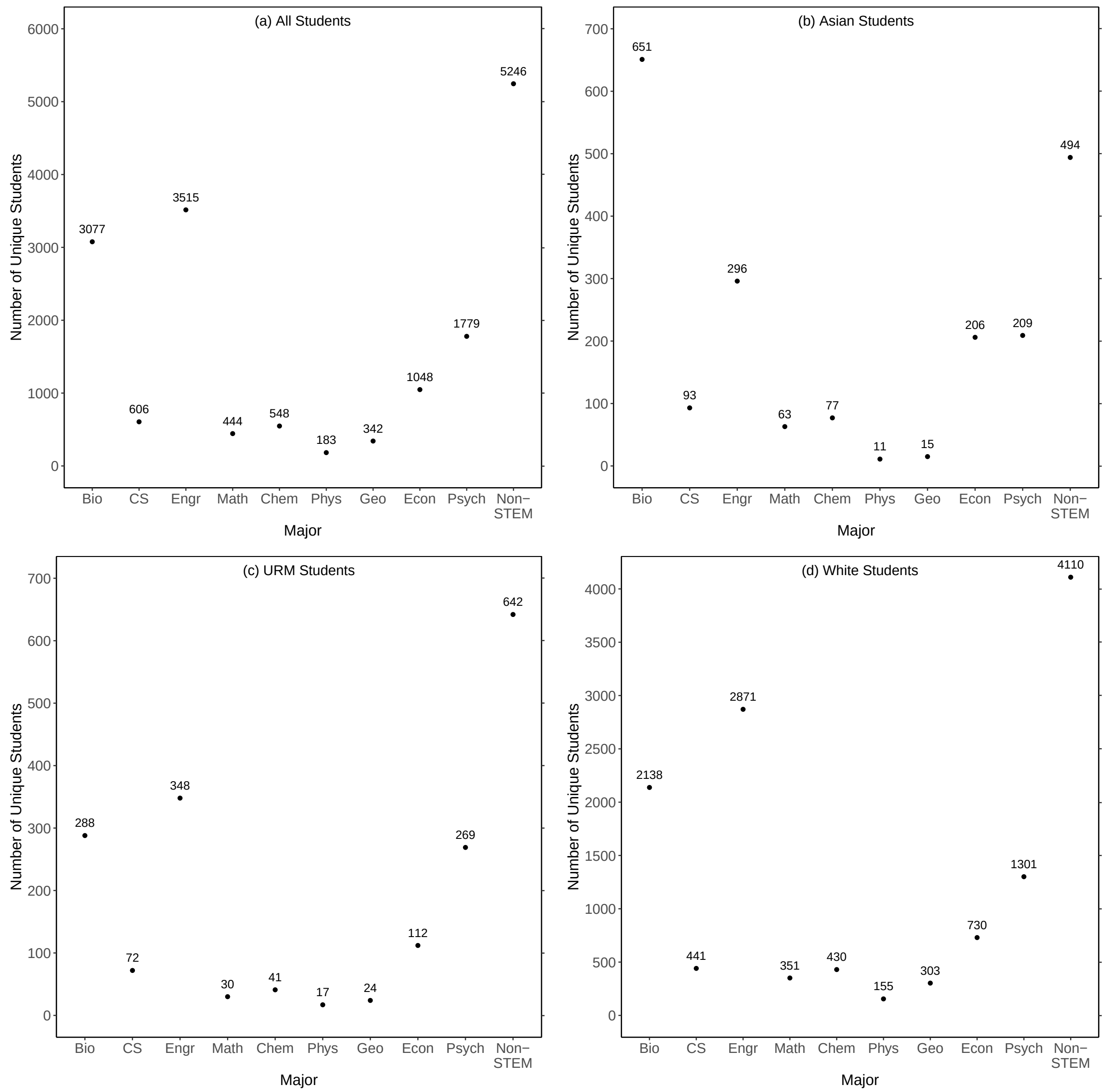


Figure 29: For each major on the horizontal axis, the number of unique students in the sample that ever declared that major is listed. Some students may contribute to the counts of more than one major (i.e., if they change majors or declare multiple majors). These counts are calculated separately for (a) all students, (b) Asian students, (c) URM students, and (d) White students. Note that the scale of the vertical axes differs for each plot.

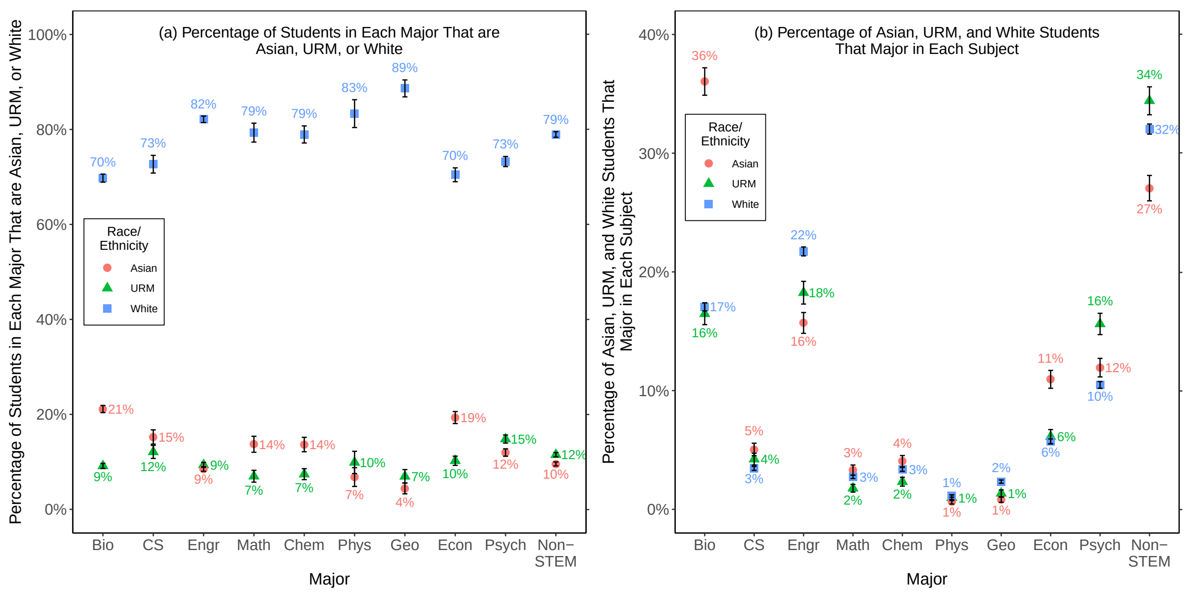


Figure 30: In (a), the percentages of students in each major that are Asian, URM, or White students are shown (i.e., the percentages in each column will sum to roughly 100%). In (b), the percentages of students in each racial/ethnic group that major in each subject are displayed (i.e., the percentages in each group will sum to roughly 100%). Discrepancies in the sum of percentages may occur due to rounding the listed percentages to the nearest integer as well as, in (b), students declaring multiple majors.

students). We can also see in Fig. 30b the differing trends for Asian students noted earlier, with Asian students more often choosing majors in the biological sciences and economics, and now we can see that conversely Asian students are less likely than URM and White students to major in engineering and non-STEM (not including economics or psychology).

Finally, another piece of information about enrollment patterns that is missing from Figs. 29 and 30 is when these students declare each major. Figure 31 shows, for each major, the average term in which students added the major as well as the peak term (that is, the term with the highest number of new students adding the major). As with Fig. 29, this is done separately for all students (Fig. 31a), Asian students (31b), URM students (31c), and White students (31d).

For the majority of majors in Fig. 31a, the peak of students adding the major is in the third term (that is, the start of their second year), with an average between the third and fourth terms. Two non-STEM majors, economics and psychology, depart slightly from this general trend, each with a peak in the fourth term and an average between the fourth and fifth terms. Two other majors, computer science and engineering, depart more significantly from the general trend in ways that can be explained by their particular implementation at the studied university.

Engineering has a peak in the first term in Fig. 31a, with an average only slightly later. Engineering is the earliest major in our data, since all admitted students who enroll in the School of Engineering are considered "undeclared engineering" majors (they have not declared subdisciplines within engineering) and so the majority of engineering students are in the engineering program in their first term. Computer science instead has the latest peak term in Fig. 31a, namely in the fifth term with a slightly later average. This is due to the structure of the computer science program at the studied university, which does not allow students to declare the major until they have completed five of the required courses for the major. These trends in engineering and computer science are important to keep in mind while considering the results presented elsewhere in this paper, since in computer science we are not able to capture attrition that occurs during the terms before a student officially declares a major. Conversely in engineering, we are able to capture almost all attrition in the first year due to the unique enrollment conditions of engineering students. This is

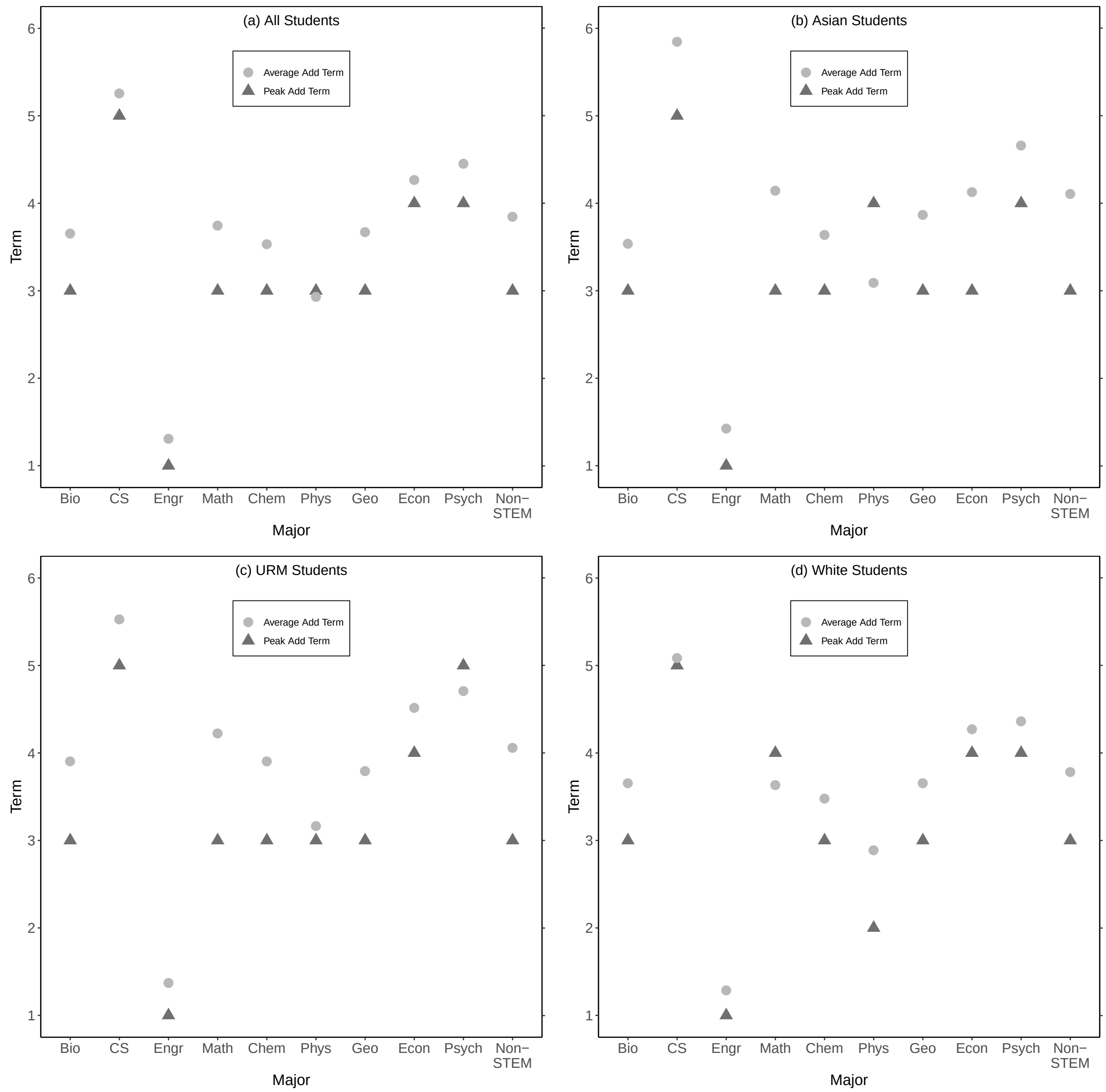


Figure 31: For each major, the term with the peak number of students adding the major in that term is plotted (triangles) as well as the average term in which students add that major (circles). This is done separately for (a) all students, (b) Asian students, (c) URM students, and (d) White students.

not possible for majors within the School of Arts and Sciences who choose majors at their convenience and, as seen in Fig. 31, not typically until at least the third term.

Turning then to Figs. 31b, 31c and 31d, we see almost identical trends as in Fig. 31a. Broadly, the differences occur for majors in which two terms have similar numbers of students declaring that major. When considering a subset of all students, which of these two terms is the peak will sometimes shift.

A more detailed accounting of the number of students that enroll in each term for each major is reported in Tables 32 and 33 in Appendix D.1. Further, summaries of the total number of unique students as well as the peak term and number of concurrent students in each major, students adding each major, and students dropping each major are available in Tables 35, 36, 37, and 38 in Appendix D.2.

9.3.2 Attrition Rates

In order to answer **RQ2**, we further considered patterns of attrition rates by race/ethnicity. In Fig. 32, we consider the drop rates of students in each major or group of majors for all students (Fig. 32a), Asian students (Fig. 32b), URM students (Fig. 32c), and White students (Fig. 32d). In Fig. 32a, we see that computer science, psychology, and other non-STEM students are the least likely to drop their major, while physics and mathematics students are the most likely to drop. We note that the relatively low drop rate of computer science majors could be due to the late declaration of the computer science major seen in Fig. 31. That is, attrition from computer science prior to when students are allowed to declare the major is not accounted for in Fig. 32.

Though the patterns in each subset of students largely mimic the pattern overall in Fig. 32a, there are a few exceptions. Figure 32e draws attention to two majors that have differential rates of attrition by race/ethnicity (namely, physics and economics). In both cases, URM students have the highest rate of attrition, followed by Asian students, and finally White students have the lowest rates of attrition.

In economics, the drop rates range from 37% of URM economics majors dropping the major (Fig. 32c) to 23% of Asian economics majors (Fig. 32b) and White economics majors

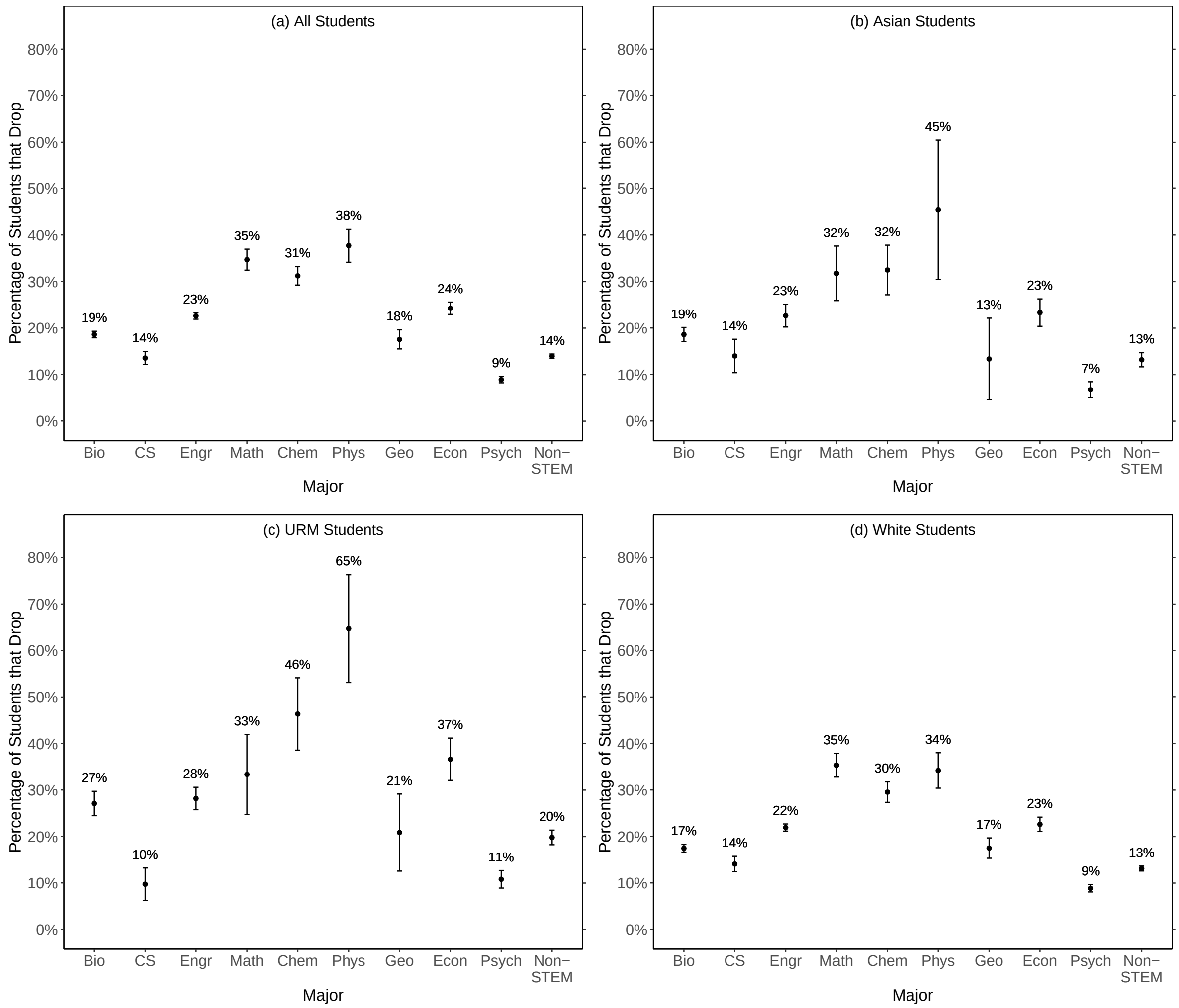


Figure 32: For each major, the percentage of students who declared the major but subsequently dropped the major is plotted along with its standard error. This is done separately for (a) all students, (b) Asian students, (c) URM students, and (d) White students.

(Fig. 32d). These differences could prove to be meaningful as more data become available, but the large standard error prohibits strong conclusions from this data. Physics, despite having even larger standard error due to a small sample size, shows an even larger disparity between URM and White students. In particular, 65% of URM physics majors (Fig. 32c), 45% of Asian physics majors (Fig. 32b), and 34% of White physics majors (Fig. 32d) drop the major. While distinguishing between the drop rates of Asian students and other students is not possible given the standard errors, the particularly high drop rate of URM physics students relative to White students is striking.

9.3.3 Trajectories of Students After Dropping a Major

Knowing now how many students drop each major, we answer **RQ3** by plotting where those dropped majors ended up in Fig. 33. In particular, the major indicated in the legends of Fig. 33a and 33b shows which major was dropped, while the plot shows which percentage of those who dropped that major ultimately earned a degree in each of the majors on the horizontal axis, including "no degree." For example in Fig. 33a, we see that among the students that drop a physics major (indicated by the line color in the legend), roughly 17% of them end up earning a degree in mathematics (by looking at this line's value above "Math" on the horizontal axis). We see that the two most common destinations for those who drop any major is either no degree or a degree in non-STEM, except for non-STEM majors who are most likely to earn no degree or a degree in psychology.

Apart from the finding that dropped STEM majors later earn degrees in non-STEM or leave the university without a degree, we see a few other interesting spikes. For instance, those who drop a physics major are likely to earn a degree in mathematics or engineering (Fig. 33a) and those who drop chemistry (Fig. 33a) are likely to earn a degree in biology. Similarly, those who drop a biological science major (Fig. 33b) are likely to earn engineering degrees. While all students who drop any major are very likely to earn no degree, the percentage of dropped majors in this category exceeds 50% for computer science (Fig. 33a), geology, psychology, and other non-STEM (Fig. 33b).

In order to further answer **RQ3**, Fig. 34 plots these same proportions of degrees earned

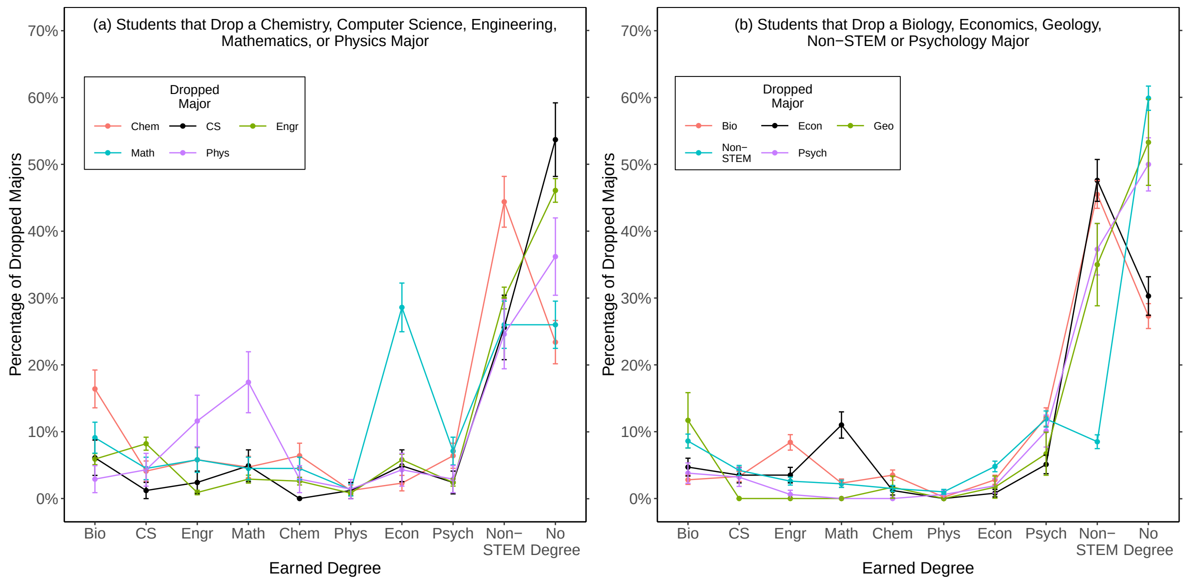


Figure 33: Among the students that drop each STEM major as well as psychology and non-STEM majors (indicated by line color in the legend), the fractions of students that go on to earn a degree in other majors, or who do not earn a degree at all, are plotted along with their standard error. Dropped majors are grouped into (a) chemistry, computer science, engineering, mathematics, and physics and astronomy majors, and (b) biological science, economics, geology, psychology, and non-STEM majors.

by students who drop a major separately for Asian students (Figs. 34a and 34b), URM students (Figs. 34c and 34d), and White students (Figs. 34e and 34f). We see for the most part very similar patterns between Asian, URM, and White students, with a few notable differences. For example, among students who drop a biology major, we see that roughly 8% of the Asian students (Fig. 34b) and 10% of the White students (Fig. 34f) go on to earn a degree in engineering, compared with roughly 2% of URM students (Fig. 34d). We see a similar pattern with a wider spread among students who drop a chemistry major, with roughly 24% of Asian students (Fig. 34a) earning a degree in biology compared with roughly 5% of URM students (Fig. 34c) and 17% of White students (Fig. 34e).

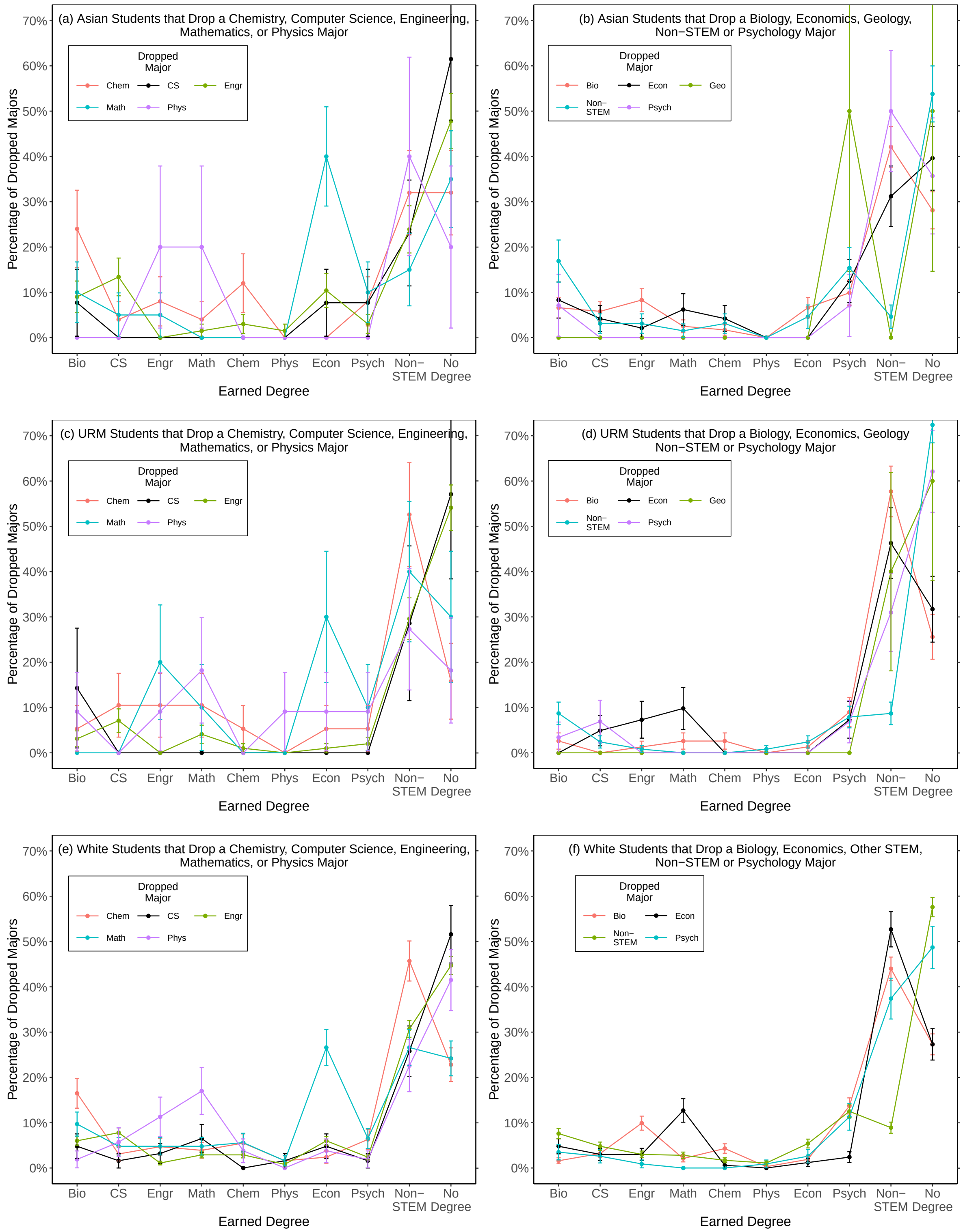
A more detailed accounting of the degrees earned by students who drop each major is provided in Tables 39, 40, 41, and 42 in Appendix D.3.

9.3.4 Degree-Earning Rates

In order to answer **RQ4**, we investigated how many students successfully earn a degree in each major. Figure 35a shows these degree-earning rates for all students in each major, while Fig. 35b shows these rates for Asian students, Fig. 35c for URM students, and Fig. 35d for White students. While these are broadly similar to a reciprocal of the drop rates in Fig. 32, since some students drop a major and subsequently declare the same major again, these degree-earning rates are a more direct measurement of persistence in a major.

Looking first at the overall rates in Fig. 35a, there are fairly wide differences across majors, from the lowest rate in physics of about 65% to the highest in psychology and other non-STEM, each at about 94%. The highest degree-earning rate in STEM occurs in computer science, with about 89% of declared computer science majors completing the degree requirements. As in Fig. 32, this can be at least partially explained by the requirements prior to declaring the major, which causes only students who have already progressed through a significant portion of the computer science curriculum to declare a computer science major.

Considering then the differences among the different racial/ethnic groups, we see relatively few differences in these degree-earning rates. As in Fig. 32, the largest difference seen here appears to be in physics, with 67% of White physics majors earning a physics degree



(Caption on next page.)

Figure 34: (Previous page.) The percentage of Asian, URM, and White students that go on to earn a degree in each major (horizontal axis) after dropping a major (indicated by line color in the legend), are plotted along with their standard error. Dropped majors are grouped into chemistry, computer science, engineering, mathematics, and physics and astronomy majors who are (a) Asian, (c) URM, and (e) White students, and biological science, economics, geology, and psychology majors who are (b) Asian, (d) URM, and (f) White students.

(Fig. 35d) compared to only 47% of URM physics majors (Fig. 35c), with Asian students very similar to White students with 64% earning a physics degree (Fig. 35b), albeit with large standard error driven by the low sample size in physics shown in Fig. 29. Similarly, White (Fig. 35d) and Asian students (Fig. 35b) are more likely to complete a degree in economics than URM students (Fig. 35c), but again the size of the standard error prevents any conclusive statements about this difference.

Across all of Fig. 35, we note that since we have combined many majors for the "non-STEM" category, this is only a measure of the number of non-STEM majors who successfully earn a degree in any non-STEM major. That is, a student who drops one non-STEM major but earns a degree in a different non-STEM major will still be counted as having successfully earned a non-STEM degree. The same is true for "engineering," which also combine several majors. The high "success rates" of computer science and psychology may be due in part to the structure of their program encouraging students to declare slightly later than other disciplines, and so this measure may not be capturing attrition that happens prior to an official declaration of major (e.g., a student intending to major in a discipline decides against it before ever declaring that major). On the other hand, since all students enrolled in the engineering school are considered "undeclared engineering" majors, the relatively low degree-earning rate of engineering reflects attrition even from the first to the second term, which is not captured for many other majors in which most students have not yet formally declared a major in their first term. Thus, each reported degree-earning rate here is a ceiling on the

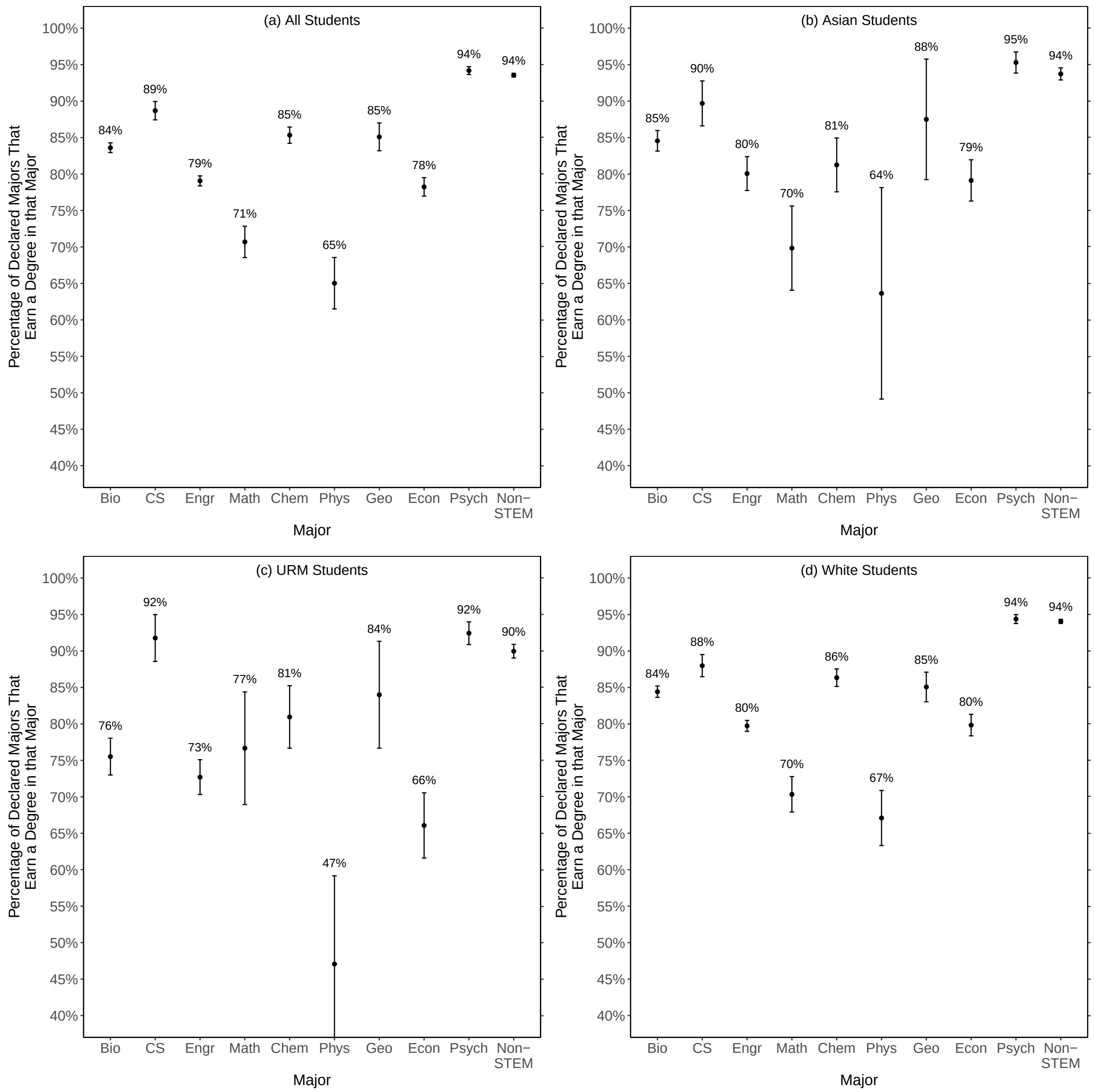


Figure 35: For each major listed on the horizontal axis, the percentages of (a) all students, (b) Asian students, (c) URM students, and (d) White students who declare that major and then earn a degree in that major are plotted along with the standard error.

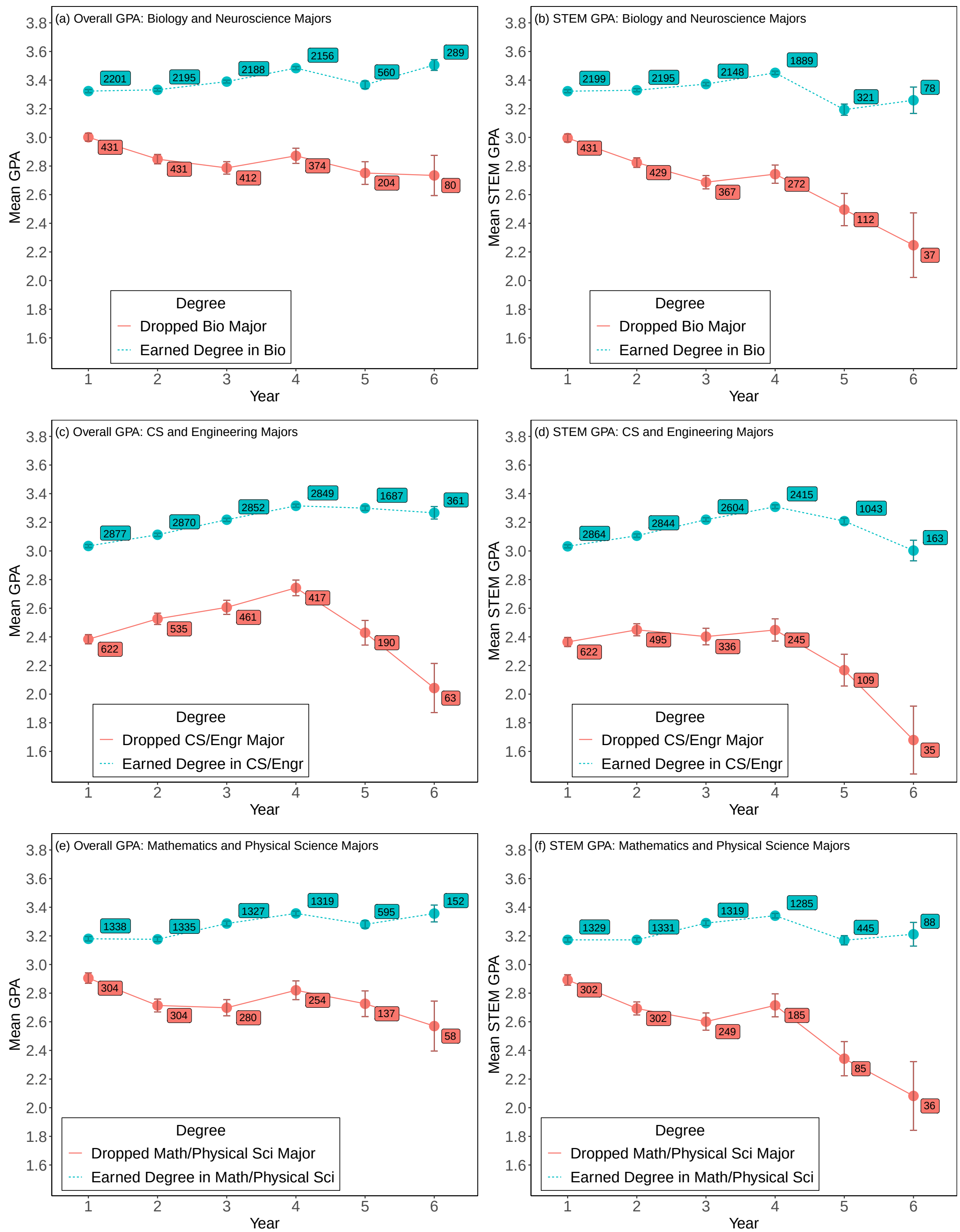
true rate that would include those students who intend to major but never declare.

9.3.5 Mean GPA of Degree-Earners vs. Major-Droppers

In order to further our understanding of why students may have dropped a given major and answer **RQ5**, Fig. 36 plots the mean GPA of STEM students who declared different sets of majors and then either earned a degree within that set of majors or dropped those majors. Note that students who dropped a major could have gone on to earn a degree with a different major or left the university without a degree. Both overall GPA (Figs. 36a, 36c, and 36e) and STEM GPA (Figs. 36b, 36d, and 36f) are plotted. Across all of Fig. 36, the large drop in sample size from year four to five and again from five to six is primarily due to students graduating. In Fig. 37, we similarly plot the mean GPA and STEM GPA of non-STEM majors who either dropped the major (i.e., left non-STEM altogether) or completed a degree with a non-STEM major.

We observe that in general, the students who drop that major have a lower GPA and STEM GPA than students who earned a degree in that major. However, the difference between the two groups varies based on which cluster of majors we consider. The GPA trends of biology and neuroscience majors (Figs. 36a and 36b) closely resemble those of mathematics and physical science majors (Figs. 36e and 36f). In both groups, those who earn the degree have a GPA and STEM GPA higher than those who drop the major by roughly 0.3 grade points higher in the first year to roughly 0.5 higher in the fourth year, with the gap widening among those who remain for five or six years. For computer science and engineering majors (Figs. 36c and 36d), those that earned a degree in the major have a GPA and STEM GPA roughly 0.6 grade points higher than those that dropped the major, consistent through the first four years of study. This difference in grade points represents a difference of one to two letter grades at the studied university, where, for example, the difference between a B and B+ is 0.25 grade points and between a B+ and A− is 0.5 grade points.

Finally, for non-STEM majors including economics and psychology (Fig. 37a and 37b), the overall GPA disparity widens over time from roughly 0.4 grade points in the first year



(Caption on next page.)

Figure 36: (Previous page.) GPA and STEM GPA over time for STEM majors. Each GPA is calculated yearly, not cumulatively. Majors are divided into three groupings: (a) and (b) biological sciences and neuroscience; (c) and (d) computer science and engineering; and (e) and (f) mathematics and physical science (chemistry, physics and astronomy, and geology and environmental science). GPA in all courses – (a), (c), and (e) – and in only STEM courses – (b), (d), and (f) – are calculated separately for two categories of students that declared at least one of the majors in each group: those that ultimately earned a degree in that group of majors and those that dropped from that group of majors. For each group, the mean GPA is plotted along with its standard error, with the sample size listed above each point and guides to the eye connecting the points.

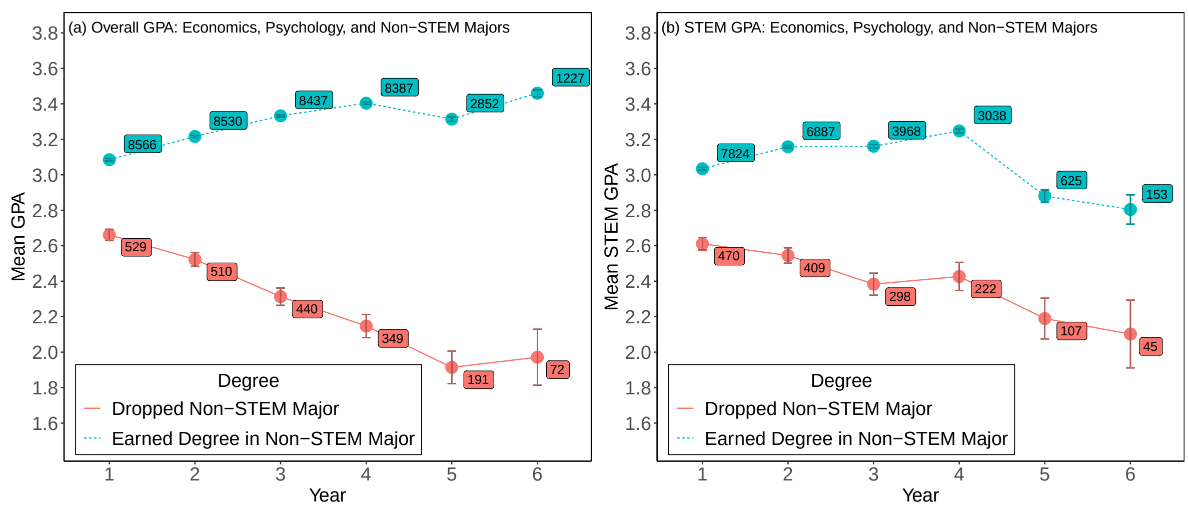


Figure 37: GPA and STEM GPA over time for non-STEM majors (economics, psychology, and other non-STEM). Each GPA is calculated yearly, not cumulatively. GPA in (a) all courses and (b) in only STEM courses are calculated separately for two categories of students that declared at least one of the majors in each group: those that ultimately earned a degree in that group of majors and those that dropped from that group of majors. For each group, the mean GPA is plotted along with its standard error, with the sample size listed above each point and guides to the eye connecting the points.

to roughly 1.2 grade points in the fourth year, while in STEM courses the GPA disparity rises from roughly 0.4 grade points in the first year to roughly 0.7 grade points in the fourth year. Notably, a much smaller fraction of students are dropping from this set – about 6% of the total number of students – which could be due in part to the wide net of considering all non-STEM majors.

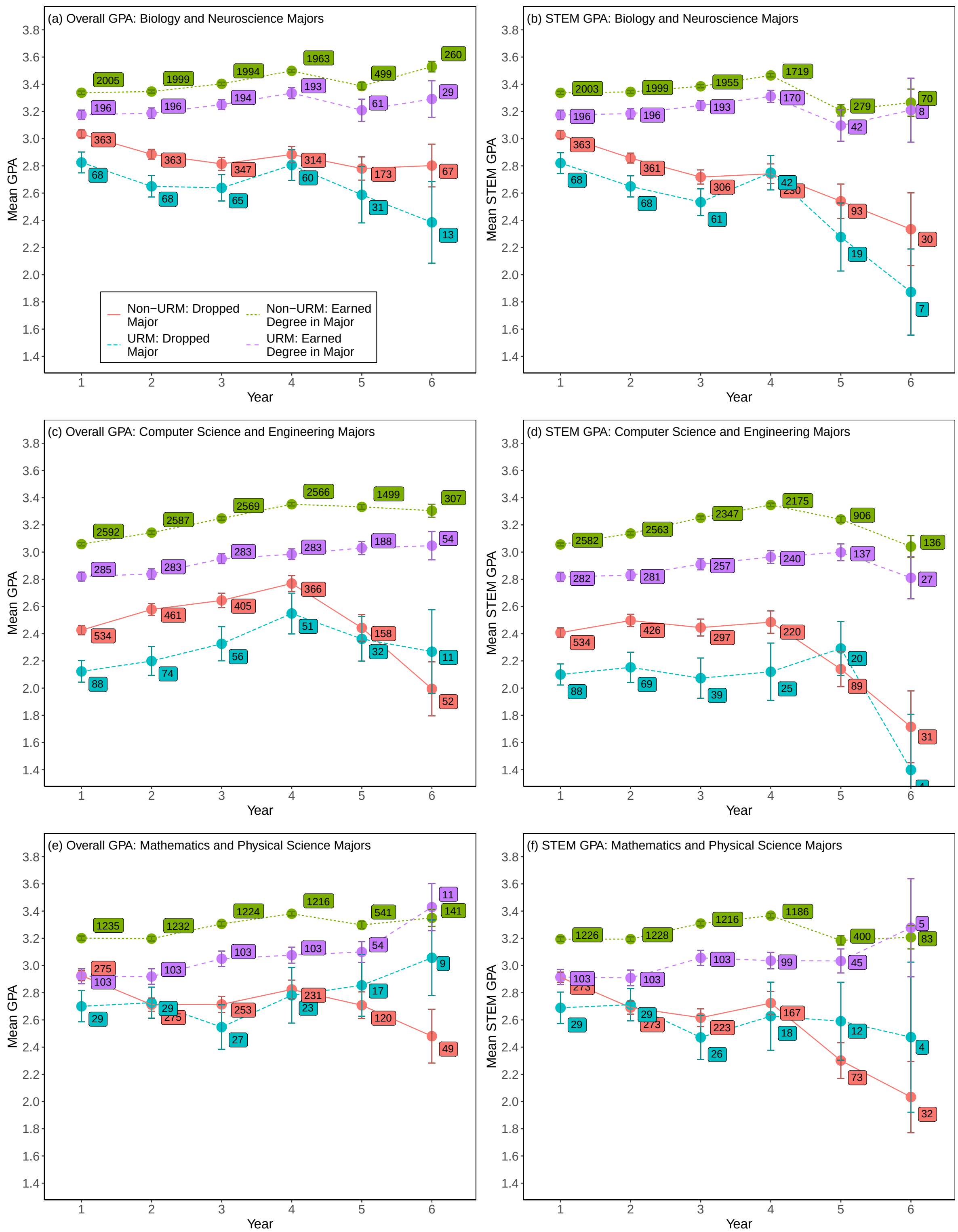
We further consider the same measures separately for URM and non-URM (i.e., Asian and White) students in Fig. 38. In this case, Asian and White students have been combined to improve clarity in the figure since there were few differences between these two groups when considering them separately. Additional versions of Fig. 38 and 39 that separately plots Asian and White students can be seen in Figs. 52 and 53 in Appendix D.4.

Figure 38 has the same subfigure structure as Fig. 36. That is, overall GPA is plotted in Figs. 38a, 38c, and 38e and STEM GPA in Figs. 38b, 38d, and 38f. Students belonging to different clusters of majors are considered separately: biological science and neuroscience majors are included in Figs. 38a and 38b; computer science and engineering majors in Figs. 38c and 38d; and mathematics and physical science (chemistry, physics and astronomy, and geology and environmental science) majors in Figs. 38e and 38f. We similarly plot the overall and STEM GPA of non-STEM majors in Fig. 39.

In all cases, the main finding is that URM students are earning lower grades on average than non-URM students, both among those who drop a given major and those who earn a degree in that major. The degree of this disparity differs little by which majors we consider, even between STEM majors (Fig. 38) and non-STEM majors (Fig. 39). In particular, URM students who earn a degree in their declared major have a GPA and STEM GPA roughly 0.2-0.3 grade points lower on average than non-URM students who earn a degree in the major, with a similar difference observed among those who drop the major.

9.4 Discussion

In this section, we first discuss general trends and then follow up with a discussion of race/ethnicity differences.



(Caption on next page.)

Figure 38: (Previous page.) GPA and STEM GPA over time for STEM majors by racial/ethnic group. Each GPA is calculated yearly, not cumulatively. Majors are divided into three groupings: (a) and (b) biological sciences and neuroscience; (c) and (d) computer science and engineering; and (e) and (f) mathematics and physical science (chemistry, physics and astronomy, and geology and environmental science). GPA in all courses – (a), (c), and (e) – and in only STEM courses – (b), (d), and (f) – are calculated separately for four categories of students that declared at least one of the majors in each group: URM and non-URM (i.e., White and Asian) students that ultimately earned a degree in that group of majors and those that dropped from that group of majors. For each group, the mean GPA is plotted along with its standard error, with the sample size listed above each point and guides to the eye connecting the points.

9.4.1 General Enrollment Patterns

Despite large differences in the numbers of students enrolling in different STEM disciplines at the studied university (Fig. 29a), there are broadly similar patterns of when those students declare the major (Fig. 31a), with some exceptions (i.e., engineering and computer science due to the constraints on when a student can declare a major). However, there are notable differences in the attrition of students from the different majors (Fig. 32a), and the corresponding degree-earning rates (Fig. 35a). Notably, a few STEM disciplines stand out as having particularly high rates of attrition (or low rates of degree completion for students who declared those majors), e.g., mathematics and physics. This is consistent with a study by Leslie *et al.* which identifies mathematics and physics as the two STEM disciplines with the highest "ability belief" (i.e., emphasis on brilliance) [158]. This trend of high attrition rates is particularly problematic for mathematics and physics, since these disciplines recruit very few students in the first place (Fig. 29). Moreover, mathematics and physics are also two disciplines with deeply hierarchical knowledge structures, which could influence student decision making, e.g., whether to leave the discipline after unsatisfactory experiences in earlier

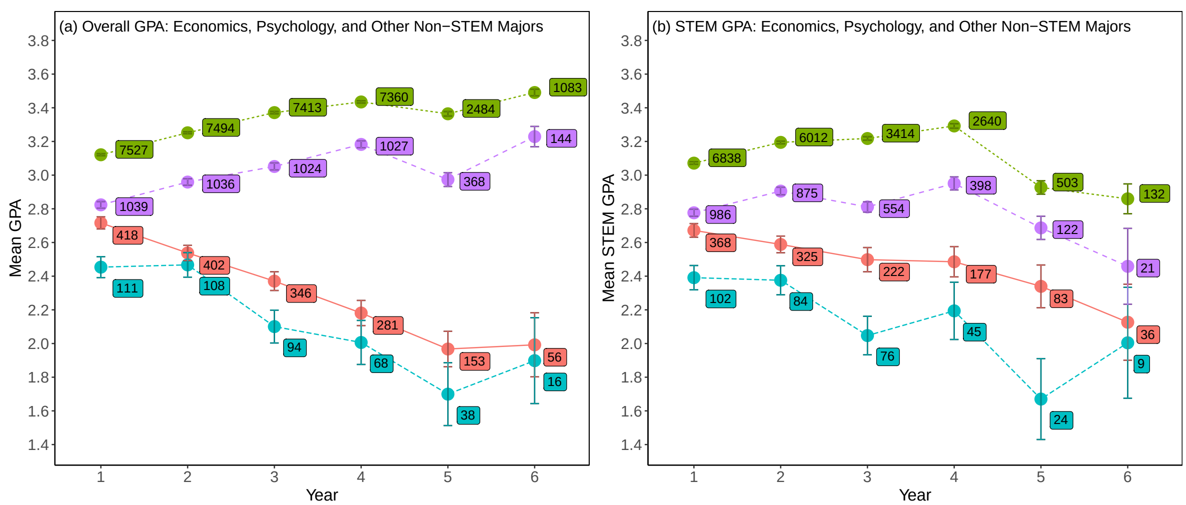


Figure 39: GPA and STEM GPA over time for non-STEM majors by racial/ethnic group. Each GPA is calculated yearly, not cumulatively. GPA in (a) all courses and (b) in only STEM courses are calculated separately for four categories of students that declared at least one of the majors in each group: URM and non-URM (i.e., White and Asian) students that ultimately earned a degree in that group of majors and those that dropped from that group of majors. For each group, the mean GPA is plotted along with its standard error, with the sample size listed above each point and guides to the eye connecting the points.

courses.

9.4.2 Enrollment Patterns by Race/Ethnicity

The most notable example of race/ethnicity differences in enrollment patterns observed in our analysis is in Fig. 30. We observe that in the biological sciences and economics, Asian students enroll at a higher rate than URM or White students. This trend is hinted at in Fig. 30a, but made clear in Fig. 30b where we see that a substantially higher percentage of Asian students choose these majors relative to URM and White students. Correspondingly, a lower percentage of Asian students choose non-STEM majors than URM and White students, and similarly fewer Asian students choose engineering majors. There are also small differences in just a few other disciplines, such as a higher percentage of URM students majoring in psychology and a higher percentage of White students majoring in engineering relative to their peers.

Apart from these differences, enrollment in the majority of majors is balanced compared to the population as a whole. This is good since it means that race/ethnicity are not playing a large role in the students' choices of major. However, it is not an equitable situation since URM students are still not being properly represented in classrooms. In particular, for URM students, the inequities often present in opportunities for college preparation and admissions are still present in the classrooms, but they are not worsened by the students' choices of major.

Moreover, we see no notable differences by race/ethnicity in when the students declare these majors (Fig. 31). Though the term in which the peak number of students adds the major may shift, the average term of adding the major is consistent among Asian, URM, and White students.

9.4.3 Attrition and Degree-Earning Rates

The primary difference in enrollment patterns found is in the attrition rates of students (Fig. 32) and correspondingly the degree-earning rates (Fig. 35). Though most majors have relatively similar drop rates and degree-earning rates among Asian, URM, and White students, a very notable exception is physics. Figure 32c shows that 65% of URM students who declare a physics major ultimately drop that major, which does not even include those students who decided to change their path before officially declaring a physics major. This is compared to 45% of Asian physics majors (32b) and 34% of White physics majors (32d) dropping the major. This is especially problematic for physics, which recruits very few students to begin with (Fig. 29), especially among Asian (Fig. 29b) and URM students (Fig. 29c). The fact that almost two-thirds of URM students drop the physics major compared to only roughly one-third of White students may be a sign of a lack of appropriate support for students and an inequitable learning environment. Further, this is consistent with prior research showing that underserved students in the sciences are especially vulnerable to stereotype threat [4, 70, 31, 46, 29, 15, 126], especially in physics, which is perceived as a field that requires a high innate "brilliance" to succeed [158].

Though no other disciplines show the same degree of racial/ethnic differences in attrition rates, Fig. 32e shows that with the exception of computer science and mathematics, URM students have slightly higher attrition rates from every major. This is a deeply troubling trend, since it runs counter to the somewhat hopeful situation noted in the balanced declarations of major (at least with regard to race/ethnicity) seen in Fig. 30. Thus, although they initially have similar aspirations as their overrepresented peers, URM students in general, including both STEM and non-STEM discplines, are being pressured to change their academic plans at higher rates than Asian and White students.

Though we had hoped to be able to shed some more light on what happens to the students who drop each major, the low sample size in Fig. 34 prevents any notable conclusions. This aspect of student attrition from various majors can be revisited as more data become available.

9.4.4 Racial/Ethnic GPA Differences

We find pervasive and troubling trends in the overall GPA and STEM GPA of STEM majors. We find in Figs. 38 and 39 that in all majors, both STEM and non-STEM, URM students consistently earn lower grades than their non-URM peers. This is true both among students that persist in their degree attainment and those that drop the major. This is problematic since these URM students are already forced to contend with stereotypes that pressure them away from STEM, and given the intricate relationship between students' expectancy or self-efficacy and academic performance [46, 292, 224, 8, 9, 10, 11, 12, 13], the feedback that these students obtain only compounds the existing societal pressures working against them and widens the racial/ethnic gap further. The reason for these troubling trends must be investigated further because they may signify lack of sufficient support, mentoring, and guidance to ensure excellence of the URM students who are already severely disadvantaged. This is especially true among the disciplines where we see high rates of attrition among URM students, most notably in physics.

Students must constantly make decisions about their academic future, and Expectancy Value Theory states that students' performance is influenced by societal stereotypes, made worse by a non-inclusive classroom learning environment in which they are underrepresented as well as previous performance feedback received. Many of these students, who come into college with a somewhat weaker preparation than their peers through no fault of their own but due to historical disadvantages, must then face these extreme pressures against their success in various STEM fields. It is critical that universities make concerted efforts to mentor and support all of their students, and improve the learning environment of these STEM disciplines to counteract the historically-rooted culture and stereotypes surrounding STEM that unfairly disadvantage URM students.

9.4.5 Limitations and Implications

One limitation of this study is the fields with the most consistent racial/ethnic differences, particularly physics, also have the lowest number of students, which limits statistical power. Also, this study limits its considerations to race/ethnicity. Other studies with larger data sets could investigate how other underserved populations such as first-generation college students or low-income students may be disadvantaged in STEM.

A critically important extension of this work would be for other institutions of different types and sizes to do similar analyses in order to broaden the wealth of knowledge available and continue to work towards the goal of equitable and inclusive education. Other institutions noting similar problematic trends can help pinpoint common sources of inequities, while institutions that do not observe these trends may be able to identify how they have structured their programs to avoid these inequitable trends. Studies such as this one can thus provide a framework for other institutions to perform similar analyses, and for particular departments to understand how their own trends differ from those of other departments at their own university.

Focus on increasing equity and inclusion in learning is especially important in the early courses for these STEM majors, since they are fraught with problematic racial/ethnic differences and may be a significant source of URM students dropping the majors. All STEM instructors should consider how they are supporting their students throughout the curriculum, since even among those that did not show differences in attrition rates we still see problematic GPA differences between URM and non-URM students. All of these issues should be addressed since they are critical for improving equity and inclusion in higher education STEM learning environments.

9.5 Acknowledgments

This research is supported by the National Science Foundation Grant DUE-1524575 and the Sloan Foundation Grant G-2018-11183.

10.0 An Example Analysis of Self-Paced Online Learning Modules

10.1 Introduction

In physics courses, we do not just want students to learn physics concepts but also to develop robust problem solving skills and be able to transfer their learning from one context to a different, new context [43, 47, 79]. Considering how to develop and evaluate instructional methods to develop and enhance students' ability to transfer their knowledge is a highly valuable research topic for STEM education. However, most existing instruments designed to assess conceptual understanding [123, 259] or problem solving skills [221, 180] cannot directly measure the ability to transfer, since there is little or no explicit learning during the test taking process, whereas the majority of research on transfer measures this ability in a controlled clinical setting.

In an earlier paper [60] we proposed a new method for measuring students' ability to transfer knowledge of physics concepts across different problem contexts by analyzing the log of clickstream data of students interacting with a sequence of online learning modules (OLMs) which contain both learning materials and assessment problems, as explained in more detail in section 10.2. We found that while introductory-level college physics students are highly capable of learning to solve a specific problem from online tutorials of rotational dynamics or conservation of angular momentum, they struggled to transfer their learning to a slightly modified problem given immediately afterwards. In a follow up study [59], we tested two different methods to enhance students' ability to transfer in an OLM sequence, and found evidence suggesting that the addition of an "on-ramp" module (a scaffolding module designed to solidify essential basic skills and concepts [190, 287] necessary for solving the remaining problems in the sequence) prior to the tutorial resulted in significant improvement in students' ability to transfer their knowledge in the rotational dynamics sequence. The current study seeks to expand upon these prior studies by investigating the mechanism by which the on-ramp module enhances transfer and how we can test the accuracy of transfer measurement using OLM sequences.

10.1.1 Measuring Transfer in an OLM Sequence

Each OLM consists of an instructional component (IC) and an assessment component (AC) which contains one or two problems. Students are required to complete at least one attempt on the AC before being allowed to study the IC, a design that draws inspiration from the frameworks of preparation for future learning [43] and constructive failure [146]. Briefly speaking, when students complete a sequence of two or more OLMs on the same topic involving similar problems, their required first attempt on the subsequent module serves as an assessment of their ability to transfer their learning from the IC of the previous module, as demonstrated in Fig. 40 adapted from [60]. When more than two modules are involved, students' performance on subsequent modules could be attributed to indirect transfer due to preparation for future learning, that is, completing the first module better prepares students to learn from the IC of the second module, which in turn results in an increase in performance on module 3.

10.1.2 Students' Different Learning Strategies and Possible Impact on Assessment

Clickstream data from an OLM sequence can be an accurate measure of students' ability to transfer on the condition that the majority of students either seriously took the required first attempt of each module or made a quick guess only when they did not know how to solve the problem. However, research on students' self-regulated learning suggests that students could adopt alternative strategies to interact with the modules. Boekaerts, for example, observed that some students may shift into a "coping mode" during learning, for which the main purpose is to preserve resources and avert damage [36].

In addition, the self-regulated learning model developed by Pintrich [223, 225] and summarized in Winne [282] suggests that students can be either achievement-oriented or goal-oriented. Achievement-oriented students focus more on and are mostly motivated by the intrinsic value of building expertise, while goal-oriented students are motivated by extrinsic value such as developing competence as a commodity that can be traded for social position.

In the case of OLM modules, the IC of each module contains detailed instructions on how to solve the specific problem in the module's AC, and only the first attempt is required in order to gain access to the IC. We expect that among students who are extrinsic value-motivated (i.e., passing the module), some of them will choose a "coping" strategy to consistently skip the first attempt (by submitting a random answer), regardless of their ability to solve the problem. This strategy avoids expending extra resources on trying to solve the problem before looking at the IC, which is only valuable for intrinsic value-motivated students who want to build mastery of the material. For students adopting this extrinsic value-motivated strategy, their transfer ability cannot be accurately reflected by their OLM performance. Therefore, those students will need to be identified and excluded from the analysis in order to avoid under estimating students' ability to transfer. However, while we exclude these extrinsic value-motivated students from analysis here, it is important to keep in mind that when implementing these self-paced online learning modules instructors should make every effort to instill an intrinsic value-motivated mindset in their students.

10.1.3 Distinguishing Between Two Different Mechanisms of the On-Ramp Module

The "on-ramp" modules in both sequences contain practice problems designed to develop and enhance proficiency of essential skills necessary for problem solving. However, if a student passed the AC on their required first attempt on the on-ramp module, they can directly move on to the next module without studying the IC. Therefore, if the on-ramp module enhances students' transfer ability by improving their proficiency in essential skills, then the benefit will not be observed for the group that passed on the first attempt (or on attempts before accessing the IC). Alternatively, if the on-ramp module mainly serves as a "reminder" for students to activate existing knowledge of essential skills, then the benefit should be more significant for those who passed on the first attempt. Distinguishing between those two mechanisms can have important implications for the future development of instructional materials to enhance students' ability to transfer.

10.1.4 Research Questions

In order to investigate these characteristics of students' engagement with self-paced online learning modules, the current study has three guiding research questions as follows.

RQ1 What fraction of students likely adopted an extrinsic value-motivated strategy when interacting with OLM sequences?

RQ2 How did data from students who adopted an extrinsic value-motivated strategy impact the accuracy of the measurement of transfer using OLM?

RQ3 Did the on-ramp module enhance students' ability to transfer by improving students' proficiency in essential skills, or by serving as a "reminder" for those who are already proficient?

10.2 Methods

10.2.1 OLM Sequence Structure

The study was conducted using online learning modules (OLMs) [56, 58, 60, 59] implemented on the Obojobo platform [20] developed by the Center for Distributed Learning at the University of Central Florida (UCF). Each OLM contains an assessment component (AC) and an instructional component (IC). Students have 5 attempts on the AC which contains 1-2 multiple-choice problems, and must make at least one attempt before being allowed to access the IC. The IC contains instructional text, figures, and/or practice questions. Specific contents of the IC used in each of the modules in the current study will be detailed in the next section. In an OLM sequence, a student must either pass or use up all five attempts on the AC before being allowed to access the next module. Students' interaction with each OLM can be divided into three stages: the pre-study (Pre) stage in which a student makes one or more attempts on the AC; the study stage in which those who failed in the Pre stage study the IC; and the post-study (Post) stage in which students make additional attempts on the AC. A student is counted as passing an AC if the student correctly answers all problems in the AC within their first 3 attempts, including both Pre and Post attempts. In other

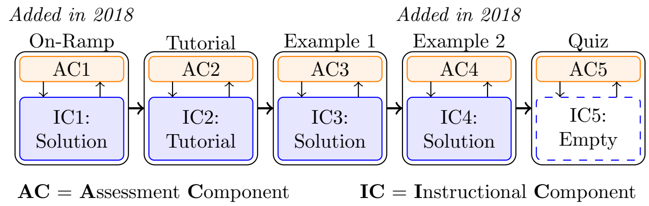


Figure 40: The sequence of Online Learning Modules (OLMs) designed for this experiment. Each OLM contains an assessment component (AC) and instructional component (IC). Students are required to make at least one attempt on the AC first, then are allowed to view the IC, and go on to make subsequent attempts on the AC. OLMs 1 and 4 were added for the 2018 implementation.

words, students who either failed on all 5 attempts or passed on their 4th or 5th attempts are considered as failing the module in the current study.

10.2.2 Study Setup

In Fall 2017, two sequences each containing 3 OLMs (specifically, OLMs 2, 3, and 5 in Fig. 40) were assigned as homework to 235 students enrolled in a calculus-based introductory physics class at UCF [60]. The 6 modules were worth 3% of the total course credit. The first OLM sequence teaches students to solve Atwood machine type problems with blocks hanging from massive pulleys using knowledge of rotational dynamics (RD). The second sequence teaches students to solve angular collision problems such as a girl jumping onto a merry-go-round using knowledge of conservation of angular momentum (AM). Both sequences are designed to develop and measure students' ability to transfer problem solving skills to slightly different contexts.

The AC of each OLM contains one problem that can be solved using the same physics principle as other ACs in the OLM sequence. The IC of OLM 2 (Fig. 40) contains an online tutorial developed by DeVore and Singh [80, 249], in the form of a sequence of practice questions. The IC of OLM 3 contains a worked solution to the AC problem, and the IC of

OLM 5 is empty since it is intended to serve the role of a quiz.

In Fall 2018, the two OLM sequences were each modified by adding two additional OLMs (shown in Fig. 40) and implemented again in the same course taught by the same instructor as homework to 241 students. Both sequences were assigned as homework, worth 3% of the total course credit in total. The first new module in each sequence is the "on-ramp" module (OLM 1 in Fig. 40), which contains an AC focusing one or more basic procedural skills necessary for solving the subsequent ACs in the OLM sequence. For the RD sequence, the on-ramp module presents students with two Atwood machine problems of the simplest form, involving one or two blocks hanging at the same radius from a single massive pulley. For the AM sequence, the on-ramp module addressed the common student difficulty of calculating both the magnitude and direction of the angular momentum of an object traveling in a straight line about a fixed point in space. The second new module in each sequence is the "Example 2" module (OLM 4 in Fig. 40), which contains in its AC a new problem that shares the same deep structure as the one in the preceding module, but differs in surface features. The IC of the module was designed in two formats: a compare-contrast format in which students were given questions that prompted them to compare the similarity and difficulty of the solutions to the problems in AC3 and AC4, and a guided tutorial format consisting of a series of tutorial-style scaffolding questions guiding them through the solution of the problem in AC4. Each form was provided to half of the student population at random. An earlier study [59] investigated the effects of those two forms of instruction and found no effect on students' behavior and performance on the subsequent module 5.

10.2.3 Data Collection and Selection

Anonymized clickstream data were collected from the Obojobo platform for all students who interacted with the OLM sequences. The following types of information were extracted from the log data following the same procedure explained in detail in Ref. [61]: the number of attempts on the AC of each module, the outcome of each attempt (pass/fail), the start time and duration of each attempt, and the start times of interaction with the IC. The duration of interaction with the IC was also extracted but was not used in the current analysis.

In addition, students' exam scores and overall course grades, each on a 0-100 scale, were also collected, de-identified, and linked to each students' log data. The exam scores consist of two midterm exams, each counting for 12% of the final course grade, and a final exam counting for 16% of the final course grade. The final course grade also contains scores from homework, lab and classroom participation.

In order to maintain a consistent sample across our analyses, only data from students who attempted every module in a sequence at least once are included. Data from seven students for the 2017 RD sequence were removed because of this reason, and two or fewer students were removed for all other OLM sequences. Data from 202 students were retained for the RD sequence 2017, 198 students in the RD sequence for 2018, 198 students for the AM sequence in 2017 and 189 students for the AM sequence in 2018.

In the Fall 2017 implementation, half of the students were given the option to skip the initial AC attempt of OLM 2 (the first OLM in that implementation) and proceed directly to the tutorial in the IC. We found in an earlier study [60] that very few students chose to exercise this option and among those who did there was no detectable impact on subsequent problem solving behavior and outcome. Therefore, in the current analysis we combined those two groups into one. Similarly, for the Fall 2018 semester we combined data from students encountering the two different versions of IC in module 4, since no difference in their behavior and outcome on module 5 could be detected [59].

10.2.4 Data Analysis

To identify and estimate the number of extrinsic value-motivated students (**RQ1**), we identified students motivated by extrinsic value by analyzing the duration of their first attempt on AC, or more specifically, the frequency of making a very brief initial attempt. As explained in the Introduction section, students who are motivated mostly by extrinsic values are more likely to consistently guess on their first attempt in order to access the instructional material.

In the current analysis, we categorize each student's first attempt as a "Brief Attempt" (BA) if the duration of the attempt is less than 35 seconds. This cutoff time is inherited from

Table 20: The number of students in each OLM sequence by their number of Brief Attempts. The BA groups consist of those who had 0-1, 2-3, or 4 Brief Attempts throughout the first four modules.

OLM Sequence	# of Brief Attempts 0-1	2-3	4
RD	100	82	16
AM	91	71	27

a careful analysis of similar OLMs in an earlier [61], and chosen as a conservative estimate for the minimum amount of time needed to read and attempt a given question. Students are categorized into four "BA groups" based on the number of BAs on the first four modules: 0-1 BAs, 2-3 BAs, and 4 BAs. Table 20 shows the number of students in each BA group for each OLM sequence. BAs on the quiz module were not considered since there was no IC for the students to access. Due to the conservative BA duration estimation, we hypothesize that the 0-1 BA group is the one with the least extrinsic value motivation, and these students are most likely to make valid first attempts on the AC.

To examine the extent to which the behavior of extrinsic value-motivated students affect the measurement of transfer (**RQ2**), we compared the Pre and Post stage pass rates of the four BA groups on all modules in the two sequences, and plotted the outcomes in Fig. 41. Following the convention established in two previous studies [60, 59], the pass rates are defined as follows. On each OLM module, the pass rates (P) of students was calculated for both the Pre-study (P_{pre}) and Post-study attempts (P_{pre}). The Pre-study pass rate on each module is calculated as

$$P_{\text{pre}} = \frac{N_{\text{pre}}}{N_{\text{total}}}, \tag{9}$$

with N_{pre} being the number of students who passed Pre-study and N_{total} being the total number of students who attempted the module. Similarly, the Post-study pass rate on each module is calculated as

$$P_{\text{post}} = \frac{N_{\text{pre}} + N_{\text{post}}}{N_{\text{total}}}, \tag{10}$$

with N_{post} being the number of students who passed Post-study (i.e., not including students who passed Pre-study). We hypothesized that the 0-1 BA group would have significantly better performance than the other two BA groups on their Pre stage attempts on modules 2, 3 and 4, because the other two BA groups are more likely to forfeit the first attempt opportunity regardless of their ability to solve the problem. We further hypothesized that the Post-study pass rates for each BA group will be very similar, since by that point every student will have either passed Pre-study or had the opportunity to learn to solve the problem in the IC. Note that students who pass the module prior to accessing the IC are still able to access the IC, but are not required to work through it before moving on to the next module.

Finally, to examine the mechanism by which the on-ramp module improves transfer of knowledge (**RQ3**), we first separate the student sample from Fall 2018 into three "on-ramp cohorts":

- **Pass On-Ramp Pre**: students who passed the on-ramp AC before accessing the IC,
- **Pass On-Ramp Post**: students who passed the on-ramp AC only after accessing the IC, and
- **Fail**: students who did not pass the on-ramp AC within 3 attempts.

Based on the analysis outcome for **RQ1** and **RQ2**, we only retained data from the 0-1 BA group for this analysis, since the performances of the other two BA groups are significantly different on certain modules, and could result in an underestimation of students' ability to transfer. The number of students in each on-ramp cohort in each OLM Sequence is listed in Table 21.

Next, we identified three comparable cohorts of students from the 2017 sample using propensity score matching, since the selected cohorts in the 2018 sample (i.e., those with 0-1 BAs) are those who engaged more seriously with the OLM sequences than the sample as a whole. In order to keep as much consistency as possible between the two samples, we only matched 2017 students with 0-1 BAs in the OLM sequence. Even though this is not a perfect comparison since there were fewer modules in the 2017 implementation, it serves to align the matching samples as closely as possible even before propensity score matching is carried out. Propensity scores were constructed using a combination of standardized scores (i.e., z-scores [98, 195, 201]) from two mid-term exams and one final exam in both semesters. Each exam is largely identical across the two semesters, with one or two questions being replaced or modified. Pass rates on all modules in both sequences are compared between the three 2018 cohorts and the three propensity score matched 2017 cohorts in order to examine the effects of the on-ramp module. These effects are considered among three groups: all students, only those who exhibited some prior mastery of the material (with the on-ramp module simply reminding them of knowledge they already possessed), and those who needed some instruction before solving the problem.

Propensity score matching was performed using R [226] and the `MatchIt` package [127]. The MatchIt algorithm retains all treated data, and attempts to find either an exact one-to-one match, or balance the overall covariant distribution for the control data.

Data analysis, statistical testing, and visual analysis were conducted using R [226] and the `tidyverse` package [279].

10.3 Results

First, we estimate the number of students that adopted an extrinsic value-motivated learning strategy (**RQ1**) by listing the number of students with 0-1, 2-3, or 4 BAs on the first four modules of each sequence in Table 20. As shown in Table 20, even with a relatively conservative criterion for classifying brief attempts, we still found that 10-15% of the students make a brief attempt on each of the four modules. On the other hand, around 50% of the students made only one or zero brief first attempts throughout the four modules.

Figure 41 shows the Pre and Post stage pass rates of students on modules 2-5, separated by the number of brief first attempts on the first four modules. Pass rates from the two sequences are plotted separately: the RD sequence in Fig. 41a and the AM sequence in Fig. 41b. In both Fig. 41a and Fig. 41b, the most prominent difference between the three BA groups is that students in the 0-1 BA group significantly outperformed the other two groups in Pre stage attempts for the Example 1 module (OLM 2, Fig 40) (Fisher's exact test on 2×3 contingency tables, $p < 0.001$ for the RD sequence and $p = 0.001$ for the AM sequence). Students in the 0-1 BA group also outperformed the 2-3 BA group on RD Tutorial Post Stage attempts ($p = 0.028$) and RD Example 1 Post stage attempts ($p = 0.018$), but the difference is not significant or the direction is reversed when comparing to the 4 BA group, which is much smaller than the other two groups.

The observation of a significant performance difference between the three BA groups on the Pre stage attempts on the Example 1 module confirmed our hypothesis (**RQ2**) that students adopting an extrinsic value-motivated strategy could have a measurable impact on the measurement of their transfer ability. Therefore, to mitigate this impact, we limit ourselves to studying the 0-1 BA group for both 2017 and 2018 student samples in the following analysis. We compared the pass rates of the 0-1 BA group from 2018 on modules 2-5 with a propensity score matched subsample in 2017 who also had 0-1 BAs on the first two modules (once again not considering the first attempt of the quiz module since there is no IC to access). The pass-rates for both sequences are shown in Fig. 42, while the p-values from Fisher's exact test comparing each pair of data points on Fig. 42 (and the upcoming Fig. 43) are listed in the first two rows of Table 22. All p-values are adjusted for Type I error due to conducting multiple tests using the Benjamini and Hochberg method [19]. The data show that there are significant performance differences in success rate between the two student populations on Tutorial Pre and Example 1 Pre attempts in the RD sequence, whereas the difference in the AM sequence is less prominent, possibly due to the success rate being very high in both samples. The differences are similar in nature but larger in magnitude compared to what was observed in a previous analysis that did not consider the different learning strategies [59]. The performance drop on Example 2 Pre attempts observed and discussed in the previous study [59] is also observed for the 0-1 BA group in the current analysis, which is primarily due to the presence of a very strong distractor option that has since been removed for future implementations.

To examine the mechanism by which the on-ramp module improves transfer of knowledge (**RQ3**), we divided the 2018 0-1 BA population into three cohorts. Since the Fail cohort is much smaller than the other two cohorts, and too small for reliable propensity score matching, we will only analyze the Pass On-Ramp Pre and Pass On-Ramp Post cohorts (see Table 21. In Fig. 43, the pass rates of the two cohorts on the same module sequence is shown side by side, compared to a propensity score matched 0-1 BA group from Fall 2017. Data from the RD sequence are shown on the top row (Fig. 43a and Fig. 43b) and the AM sequence in the bottom row (Fig. 43c and Fig. 43d). The adjusted p-values of Fisher's exact test between each pair of points are listed in the last four rows of Table 22.

It can be clearly seen from Fig. 43 that the Pass On-Ramp Pre cohort is responsible for the majority of the differences on Pre-study attempts between the 2017 and 2018 samples. For the RD sequence, the differences are no longer statistically significant for the Pass On-Ramp Post cohort after p-value adjustment [19]. For the AM sequence, none of the differences were statistically significant after p-value adjustment [19] for either the Pass On-Ramp Pre or Pass On-Ramp Post cohorts.

10.4 Discussion

We found that roughly half of the students either occasionally or frequently adopted a learning strategy that is likely to be motivated by extrinsic value: making abnormally short attempts on their required first attempts on some or all of the first four modules. These brief attempts could have been generated by students who were either guessing or copying the answer from a peer. While an occasional brief attempt may indicate a lack of confidence in one's own knowledge, continuous brief attempts on multiple modules are much more likely a strategic choice to save time on the task, especially given the significantly higher correct percentage on attempts after studying the IC. This strategy fits well with Boekaerts's description of students being in a "coping mode," in which their goal is to pass the module while preserving resources and averting "damage" as much as possible [36]. According Pintrich's self-regulated learning model [223, 225], this strategy is motivated less by the intrinsic value of acquiring problem solving expertise and more by the extrinsic value of acquiring the points for passing the module.

For students who adopted such a "coping" strategy, their transfer ability can no longer be correctly measured using OLMs, as their brief Pre-study attempt on the following modules do not always reflect their true ability to transfer their learning from the current module.

Our analysis suggests that data from those students in a previous study resulted in an underestimation of students' ability to transfer knowledge from the Tutorial module (module 2) to the Example 1 module (module 3), although most of the qualitative conclusions remain the same.

It is worth mentioning that an alternative explanation of our observation is that students who frequently adopt the "coping" strategy have a lower level of overall mastery of the subject. Therefore, they would not have been able to pass the required Pre stage attempt even if they had spent more time on it. Thus, including those students would not result in an underestimation of students' transfer ability. However, while this may be true for some students, we do not think that it applies to the majority of students in the 2-3 BA and 4 BA groups. This is because their performance on modules 2, 4 and 5, as well as on the Post stage of the module 3 are identical to that of the 0-1 BA group, which suggests that their overall abilities are similar and therefore that the difference observed in the Pre stage attempts on module 3 is mostly due to difference in strategical choice.

Another major finding of the current analysis is that the benefit of the on-ramp module in facilitating transfer (as detected during Pre stage attempts of subsequent modules) predominantly occurs among students who can pass the on-ramp module before accessing the instructional component. The difference is much more prominent for the more challenging RD sequence, and less so for the easier AM sequence. This unexpected observation holds true even after we used propensity score matching between the two semesters to control for the fact that the Pass On-Ramp Pre cohort likely includes more students who began the OLM sequence with a better understanding of the underlying physics concepts, or who are more motivated to study the subject, than the Pass On-Ramp Post cohort. A possible explanation can be obtained using the basic principles of information processing theory [257, 246]. For students who already possess the essential skills, attempting the on-ramp module assessment prompted them to retrieve those skills from long-term memory and import them into working memory. All or part of those skills remained either in the working memory or in a highly active state when the students moved on to the subsequent modules, thereby freeing up cognitive capacity for those students to better comprehend the additional complexity of the Tutorial and Example 1 modules. On the other hand, for those who had not yet mastered the essential skills, the IC of the on-ramp module was sufficient for them to learn the skills, but not enough for them to achieve a higher level of proficiency. Therefore, activating those newly learned skills on the subsequent modules required a significant amount of cognitive load, limiting students' abilities to consider additional complexity.

A straightforward and testable implication of this explanation is that more practice problems on those essential skills will increase students' ability to learn and transfer on subsequent modules. In addition, it may be beneficial to distribute those practices rather than clustering them immediately prior to the tutorial sequence, as distributed practice has been shown to be beneficial to skill acquisition and recall [85, 120].

It must be pointed out that our use of propensity score matching to control for the fact that students in the Pass On-Ramp Pre cohort have better incoming knowledge than those in the Pass On-Ramp Post cohort is far from perfect. This is because there may not be a strong correlation between students' level of mastery on rotational dynamics and their overall exam performance. A more accurate propensity score can be constructed if additional modules on the same topic are assigned to students prior to the tutorial sequence. Such modules have been created and administered in the Fall 2019 semester, enabling more accurate analysis to be conducted in the future.

10.4.1 Implications for Online Education Research

Our analysis shows that students' behavior in a self-regulated online learning environment frequently deviates from what was intended or expected by the instructor due to differences in student motivation and learning strategies. Those behaviors, such as frequently guessing on problems, can have a substantial impact on the outcomes of data analysis if not properly accounted for. While this type of behavioral diversity certainly adds to the complexity of data analysis, we argue that it should not be seen as a downside of conducting educational studies in an online environment. On the contrary, the ability to detect the presence of different behaviors, quantify their frequencies, and account for them during data analysis is a unique strength of online educational research, since these extrinsic value-motivated strategies are certainly not exclusive to online learning. Furthermore, our recent study [109] shows that similar brief-attempt behavior can be observed even when students are in a proctored clinical environment. Therefore, analyzing online learning data provides a powerful method to identify these strategies and reduce their impact on the analysis results.

The advantage of online learning research in providing detailed information on students' learning behavior also enabled us to reveal the mechanism by which the "on-ramp" modules enhance students' ability to transfer, which is quite different from what we had expected. Our earlier analysis [59] on the two different versions of module 4's IC demonstrated that results expected based upon prior theoretical frameworks (i.e., the benefit of compare-contrast tasks in instruction) may not actually work in reality. Further, the current analysis of the module sequence shows that even when an intervention works (i.e., the inclusion of the on-ramp module increasing subsequent performance), it may work for a very different reason than expected, which serves as another testimony to the highly complex nature of teaching and learning. More importantly, these results signify the critical importance of discipline-based educational researchers, as well as the challenges they face as "Education Engineers" in taking the general principles proposed in education literature and trying to build concrete interventions to actually improve student learning in STEM disciplines.

Last but not least, the current study is an exploratory attempt at evaluating the effectiveness of instructional materials by comparing the outcomes of students enrolled in two consecutive semesters, and controlling for the extrinsic variances using propensity score matching. Compared to the more common method of conducting randomized AB experiments [57, 55] (i.e., an experiment where the sample is randomly assigned to either group "A" or "B" and assigned either the treatment or control condition, then subsequently reversed), the current method is significantly easier to implement in real classroom settings and introduces very little disruption into students' learning process. In addition, this method allows for a much larger sample size since each group is an entire class rather than a fraction of a class. While it introduces significantly more variance between different semesters must be controlled for, the current method can be particularly valuable when randomized trials are difficult to implement, such as during the COVID-19 outbreak, which presents significant hurdles for recruiting human subjects, while students are already subjected to great stress trying to cope with the challenges of remote learning.

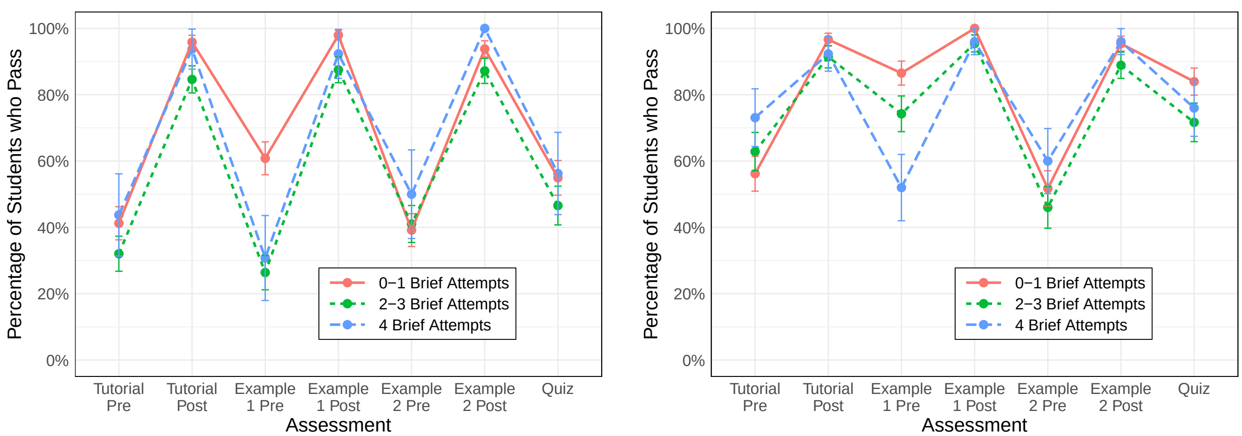
Finally, this method of administering self-paced learning tools and the ability to conduct analyses using clickstream data of students' interactions with these tools provides a valuable opportunity for future research in gender equity and students motivational factors in physics. In an extension to the work done in previous chapters which operated on a curricular or course grade scale, the study presented here provides a blueprint for future studies to investigate the much smaller scale of individual assignments.

10.5 Acknowledgments

The authors thank the Learning Systems and Technology team at UCF for developing the Obojobo platform. This research is partly supported by NSF Grant DUE-1845436.

Table 21: The number of students in each OLM sequence that fall into each on-ramp cohort among those with 0-1 BAs. The cohorts consist of those who passed during on-ramp Pre-study attempts (“Pass On-Ramp Pre”), those who passed during on-ramp Post-study attempts (“Pass On-Ramp Post”), and those that failed the on-ramp assessment (“Fail”). Since the on-ramp module was only included in Fall 2018, only students from 2018 are included here.

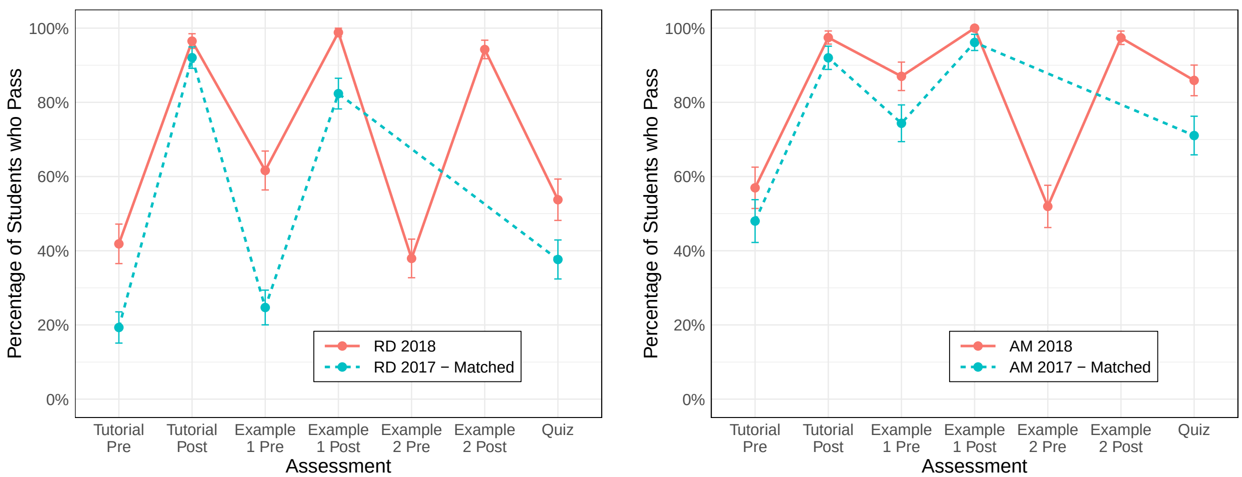
OLM Sequence	Pass On-Ramp Pre	Pass On-Ramp Post	Fail
RD	32	57	11
AM	32	47	12



(a) Rotational Dynamics

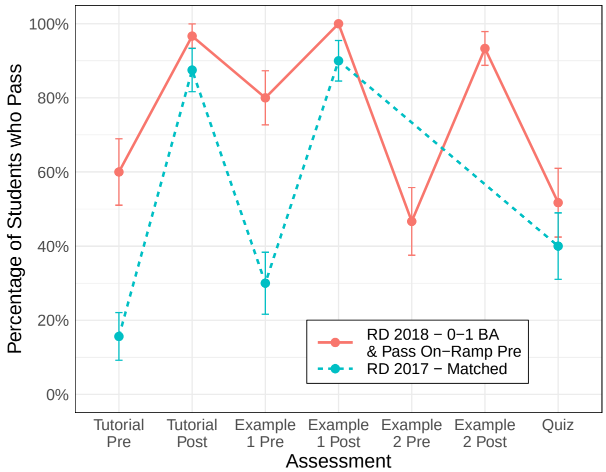
(b) Angular Momentum

Figure 41: Students are grouped by their number of Brief Attempts throughout the OLM sequences for (a) Rotational Dynamics and (b) Angular Momentum. The pass rates of these groups in each module along with their standard error are plotted.

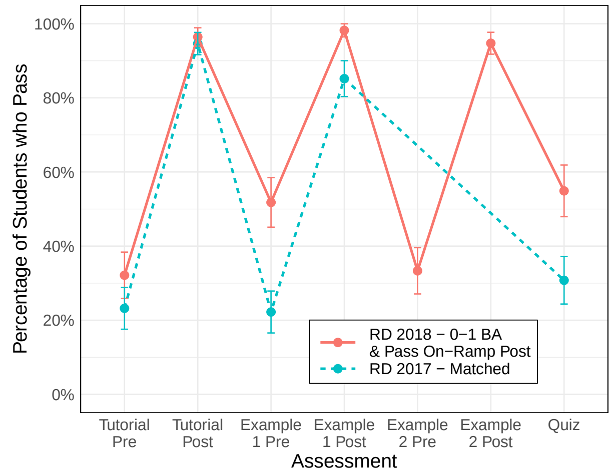


(a) Rotational Dynamics (b) Angular Momentum

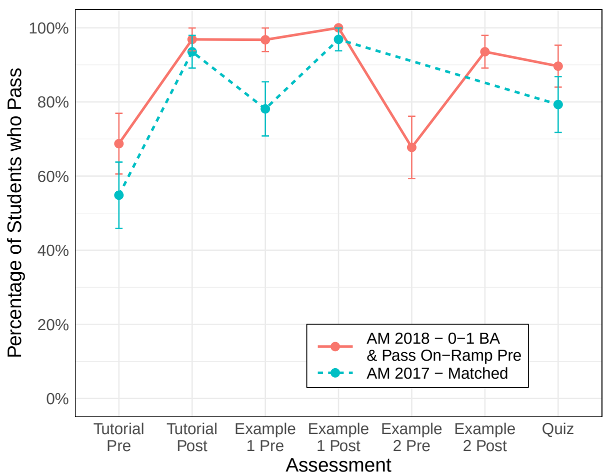
Figure 42: Using propensity score matching on course exam scores, a subset of 2017 students is matched to 2018 students with 0-1 Brief Attempts. The pass rates of these two samples are then plotted separately for (a) Rotational Dynamics (RD) and (b) Angular Momentum (AM).



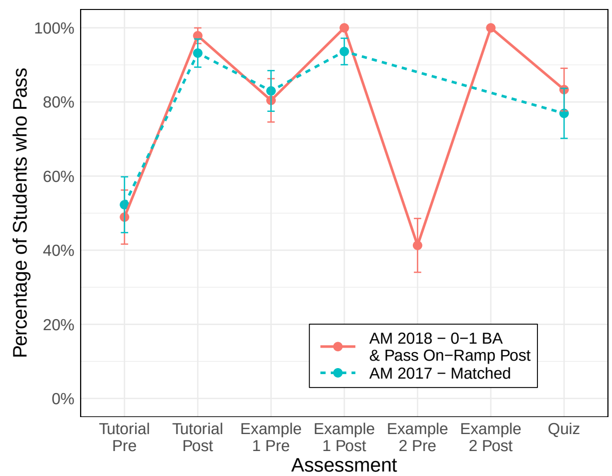
(a) Rotational Dynamics: Matching 2018 Pass On-Ramp Pre students.



(b) Rotational Dynamics: Matching 2018 Pass On-Ramp Post students.



(c) Angular Momentum: Matching 2018 Pass On-Ramp Pre students.



(d) Angular Momentum: Matching 2018 Pass On-Ramp Post students.

Figure 43: Using propensity score matching on course exam scores, a subset of 2017 students is matched to 2018 students with 0-1 Brief Attempts in either the (a) and (c) Pass On-Ramp Pre or (b) and (d) Pass On-Ramp Post cohorts. The pass rates of these two cohorts are plotted separately for (a) and (b) Rotational Dynamics (RD) and (c) and (d) Angular Momentum (AM).

Table 22: A list of p-values from Fisher's exact test comparing the performance of 2018 students and matched 2017 students on each common assessment in the listed figure. The p-values have been adjusted using the Benjamini & Hochberg method [19].

Fig.	Tutorial Pre	Tutorial Post	Example 1 Pre	Example 1 Post	Quiz
42a	0.003	0.330	< 0.001	< 0.001	0.054
42b	0.333	0.265	0.166	0.306	0.166
43a	0.001	1.000	0.001	0.395	0.438
43b	0.498	1.000	0.008	0.028	0.028
43c	0.764	0.766	0.267	1.000	0.766
43d	0.835	0.835	0.835	0.835	0.835

11.0 Conclusion and Future Outlook

These various studies paint a picture full of various and sundry ways that inequities pervade an education in physics and how extensive institutional data can be an invaluable resource in studying these inequities. The study of the grades earned by physics majors discussed in Chapter 2 reveals that gender disparities in physics grades do not appear in advanced courses, but are instead concentrated in the introductory courses where male physics majors earn higher grades than female physics majors. Notably, these inequitable grade differences occur during the first year of study when students are making many important decisions about their academic trajectory. These trends, which we detect among only those who persisted in physics long enough to declare the major, provide hints of why other prospective physics majors chose to take another path. This study can be extended by gathering additional information that can be used to identify those students who choose to change away from a physics major in order to identify ways in which these prospective physics majors may be better supported in their pursuit of an education in physics.

Following the investigation of physics grades in Chapter 2, the study presented in Chapter 3 shifts the focus to the motivational characteristics of physics majors. We find that physics majors' perceived recognition and interest are consistent across all levels of their undergraduate education and among first-year graduate students, while self-efficacy and physics identity show a decline as students progress through the curriculum. This decline begins as early as the first to the second year, once again identifying a troubling trend that coincides with the time that students are deciding whether or not to continue pursuing a major in physics. Currently, the statistical power of the motivational data coupled with the already low representation of women and underrepresented minority students limits comparisons by gender or race/ethnicity. As data continues to accumulate, future studies could extend this work by considering how these trends differ by gender or race/ethnicity.

Following these studies that focus on the experiences of physics majors, we broaden our scope in order to provide important context to the inequities observed in introductory physics. The students in introductory physics courses are primarily first-year engineering majors, who take physics as part of their common first-year curriculum. The studies presented in Chapters 4, 5, and 6 focus on these engineering students in order to investigate how their experiences in physics differ from their experiences in other disciplines. We find in Chapters 4 and 5 that for these engineering students, physics plays a supporting role in their success in future engineering courses, which is primarily dependent on their advanced mathematics courses. However, physics is the only discipline among all of their coursework (including mathematics, chemistry, and engineering courses) in which gender differences are observed. Chapter 5 shows that among engineering majors, women perform better on average than men in their mathematics, chemistry, and engineering courses – a trend consistent with their high school GPA – while the opposite is observed in physics, with men earning higher grades than women. Chapter 6 explores this further, wherein we find that physics is also the discipline with the largest gender difference in self-efficacy among engineering majors, again with men having higher self-efficacy than women. These gender differences in introductory physics grades and self-efficacy in physics could be an explanation for the severe underrepresentation of women in "physics-heavy" engineering programs, and these studies could be extended by surveying or interviewing engineering students in order to learn more about their motivations for choosing their engineering program and why they did not choose other available programs.

Extending the context of our investigations of physics further, Chapters 7, 8, and 9 examine how enrollment patterns in physics compare to those of other STEM disciplines. Amidst the backdrop of broader findings that compare the enrollment patterns and grades earned by STEM majors among various demographic groups (considering gender, race/ethnicity, first-generation status, and low-income status), a consistent pattern emerges among each of these three chapters about physics. Regardless of the demographic characteristics considered, physics consistently has the lowest retention rates among STEM majors. In addition to the generally low number of students to begin with, especially women and underrepresented minority students, physics fails to retain these students. The previous chapters found many and various ways in which the environment in physics was inhospitable to underrepresented students, and these studies now provide a full context that this situation is not a broader trend in STEM but rather a uniquely inequitable situation in physics.

Finally, we return to the physics classroom in Chapter 10, which presents a study on how extensive data logs from an online homework system can be used to conduct studies that probe student problem-solving and learning more directly than the studies of students' course grades. This chapter presents a blueprint for future studies that can be used in conjunction with responses to motivational surveys in order to glimpse how students with, for example, higher or lower self-efficacy engage differently with problem-solving and instructional material.

With the research presented throughout this thesis, we have identified numerous areas in which physics produces inequitable outcomes for students, further perpetuating an environment that is not inclusive to students. These results should be used as a baseline so that future studies which incorporate strategies to improve the environment in physics can detect whether or not these strategies succeeded in improving the inclusivity of physics on a scale that is undetectable by course grades alone.

Appendix A Supplementary Material: Physics SEM

A.1 Grade Distributions by Gender

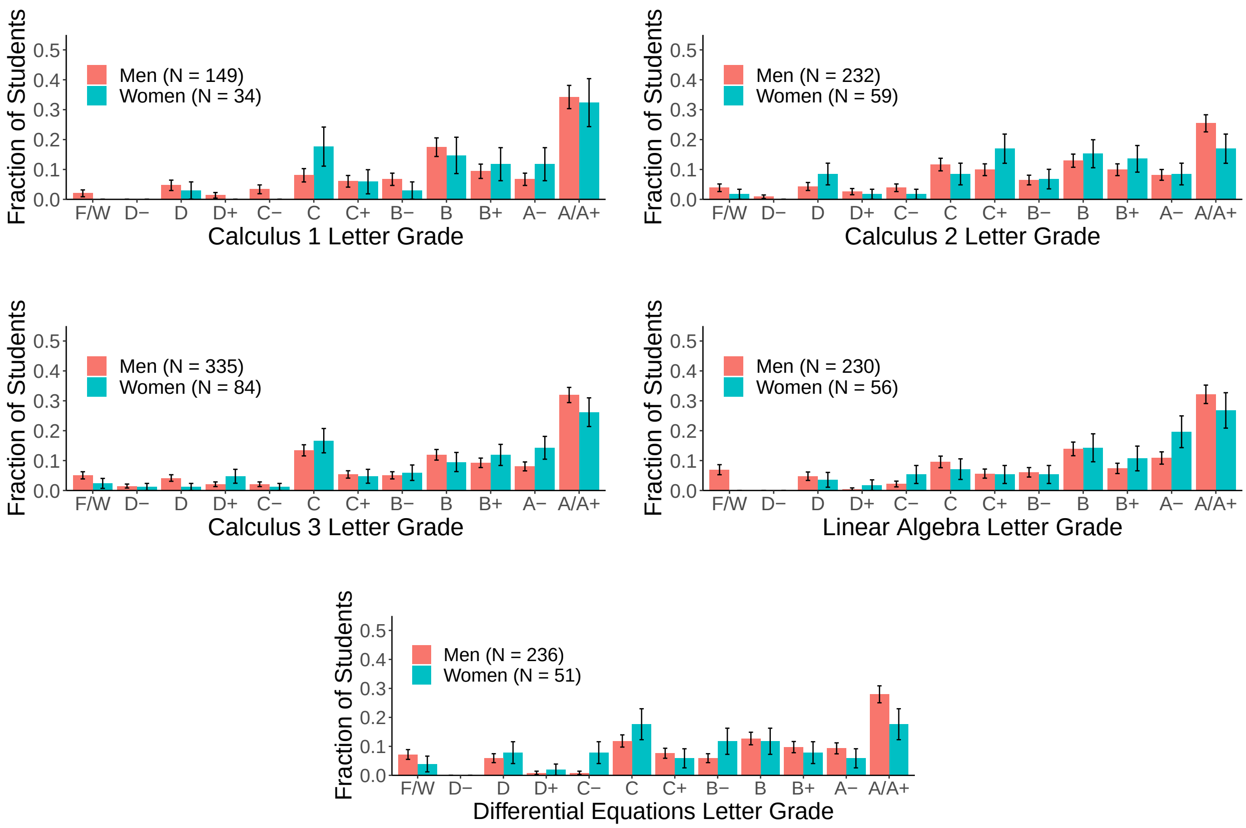


Figure 44: Grade distributions of physics majors in required mathematics lecture courses, plotted separately for men and women. The proportion of each gender group that earns each letter grade is plotted along with the standard error of a proportion.

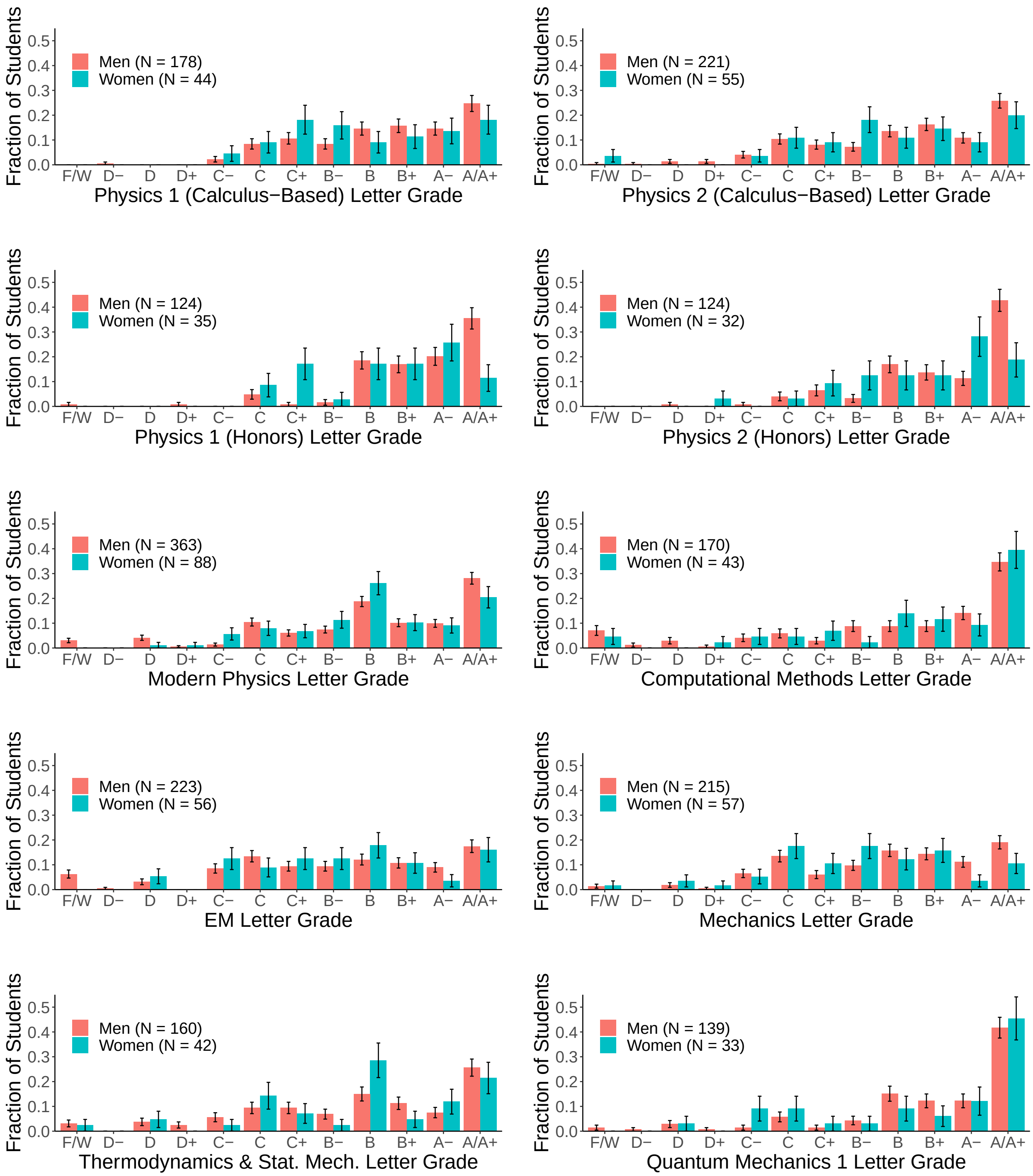


Figure 45: Grade distributions of physics majors in required physics lecture courses, plotted separately for men and women. The proportion of each gender group that earns each letter grade is plotted along with the standard error of a proportion.

Appendix B Supplementary Material: Self-Efficacy and Performance of Engineering Students

B.1 Self-Efficacy Distributions

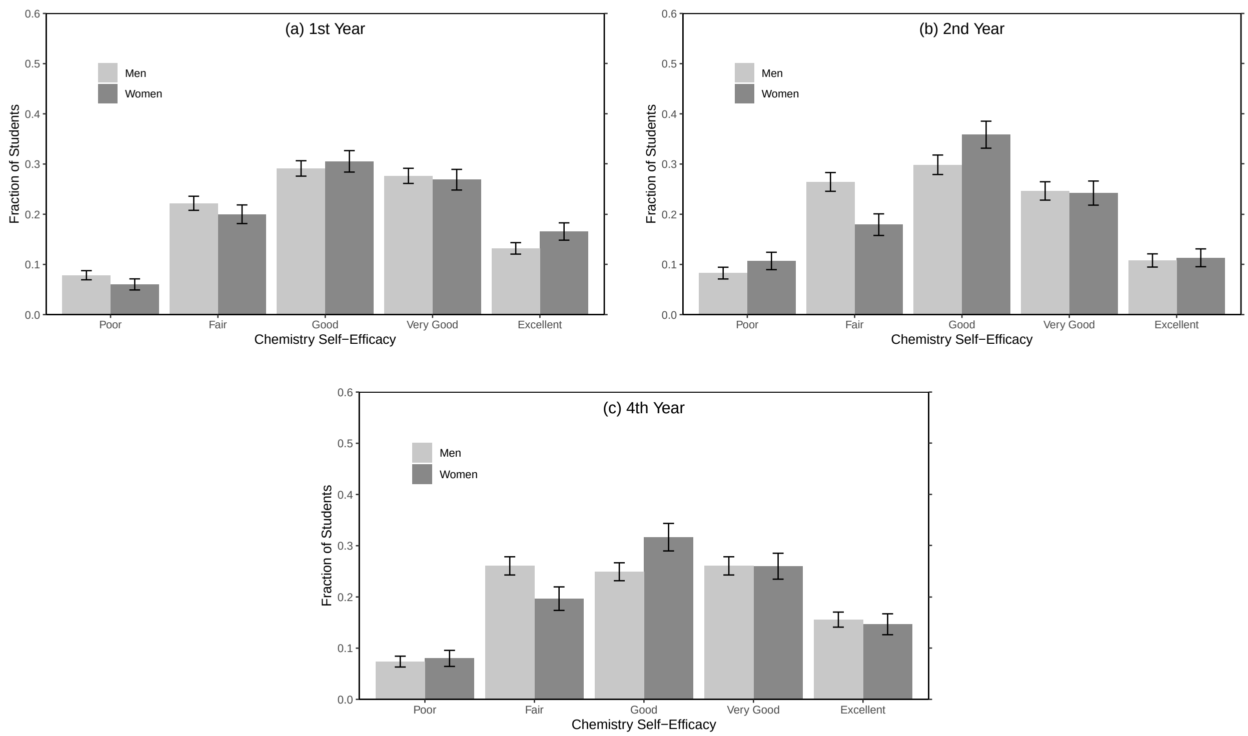


Figure 46: Distribution of responses to the chemistry self-efficacy prompts the surveys taken by students in their (a) 1st year, (b) 2nd year, and (c) 3rd year. The fraction and standard error of men and women, respectively, who answered each of the five options are plotted.

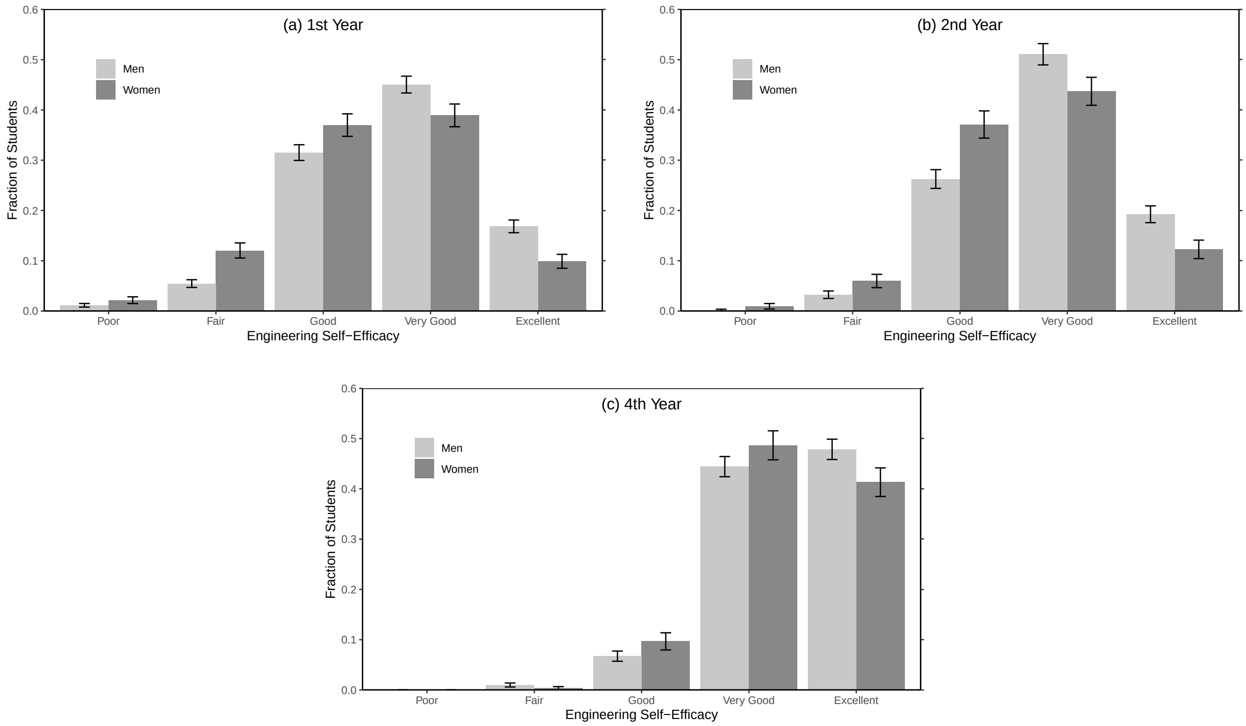


Figure 47: Distribution of responses to the engineering self-efficacy prompts the surveys taken by students in their (a) 1st year, (b) 2nd year, and (c) 3rd year. The fraction and standard error of men and women, respectively, who answered each of the five options are plotted.

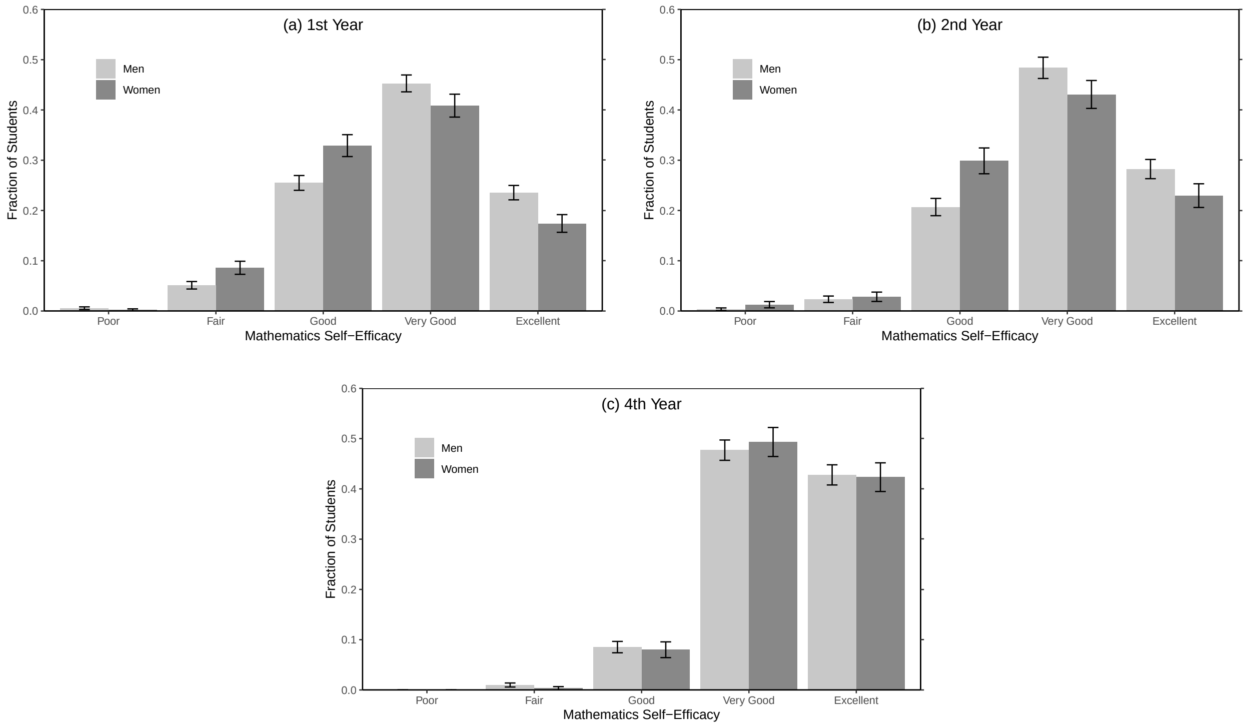


Figure 48: Distribution of responses to the mathematics self-efficacy prompts the surveys taken by students in their (a) 1st year, (b) 2nd year, and (c) 3rd year. The fraction and standard error of men and women, respectively, who answered each of the five options are plotted.

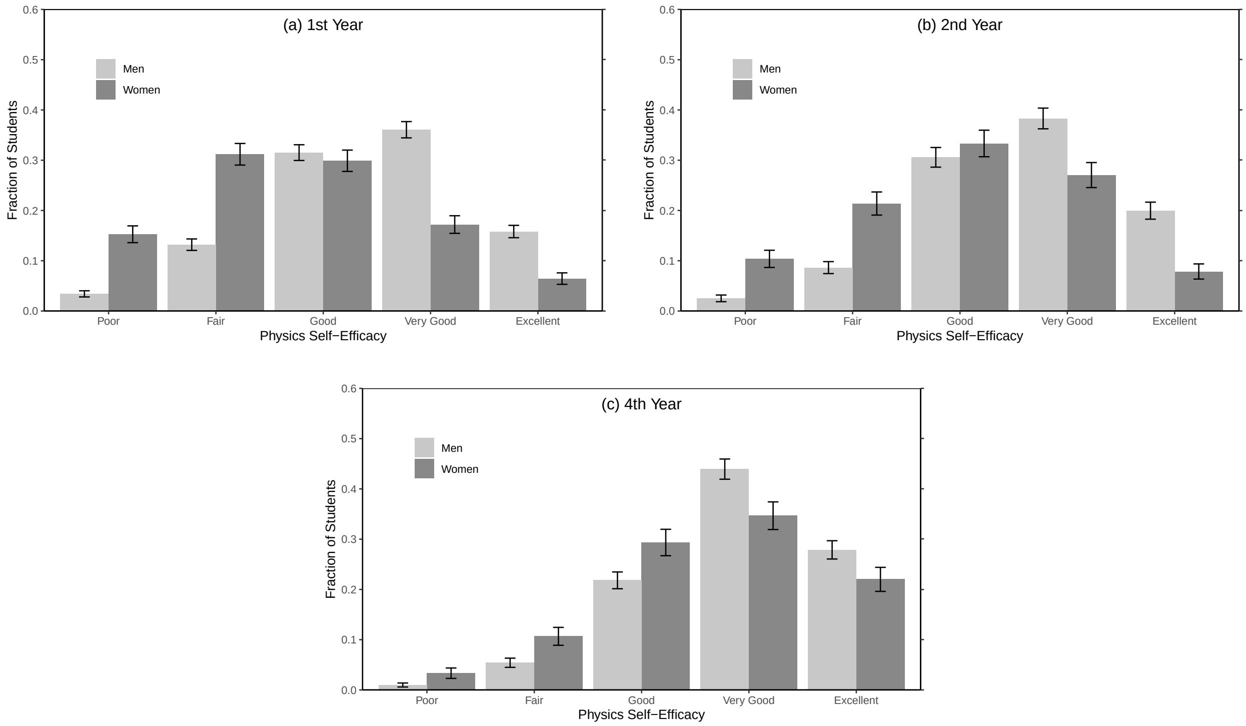
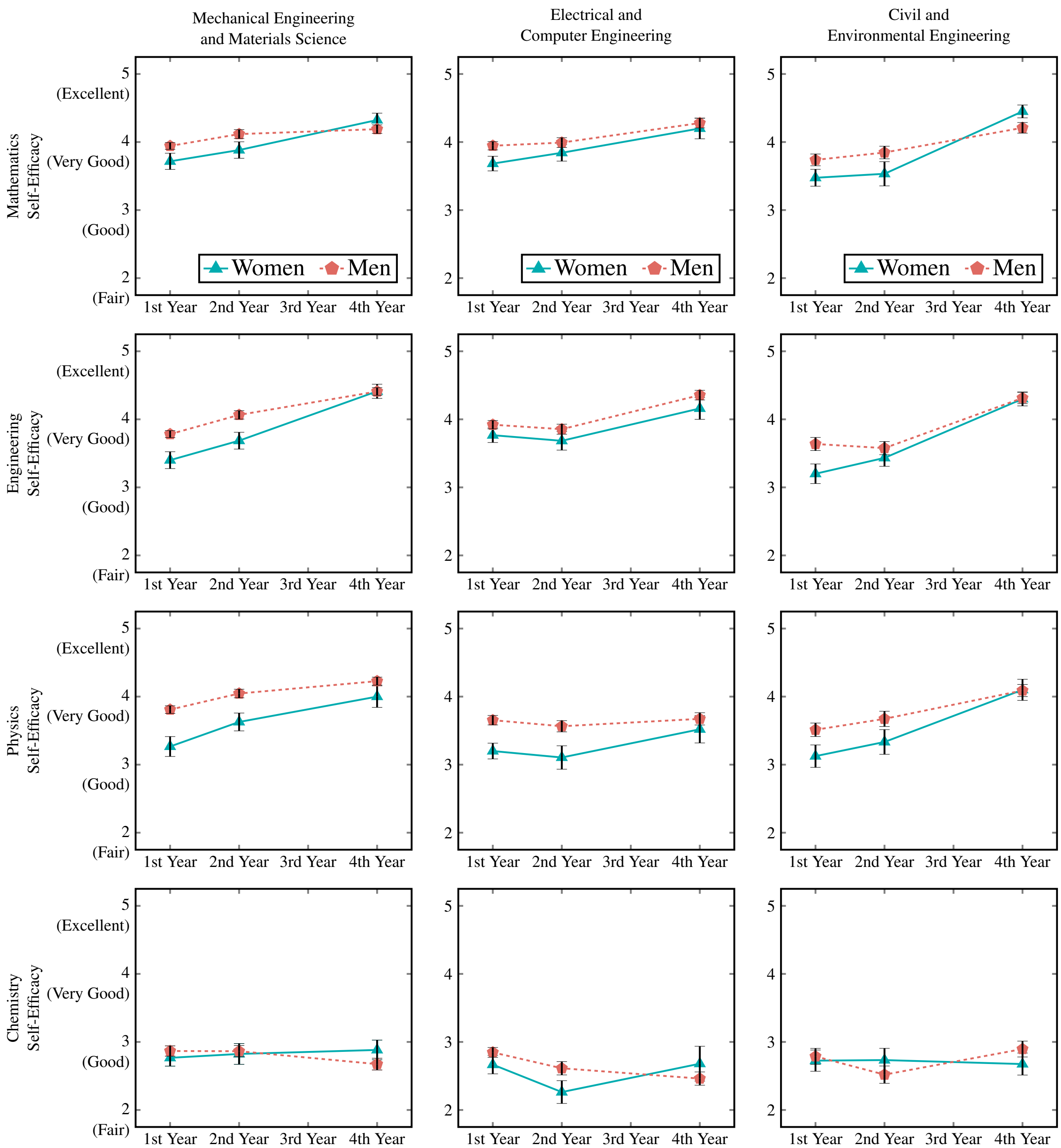


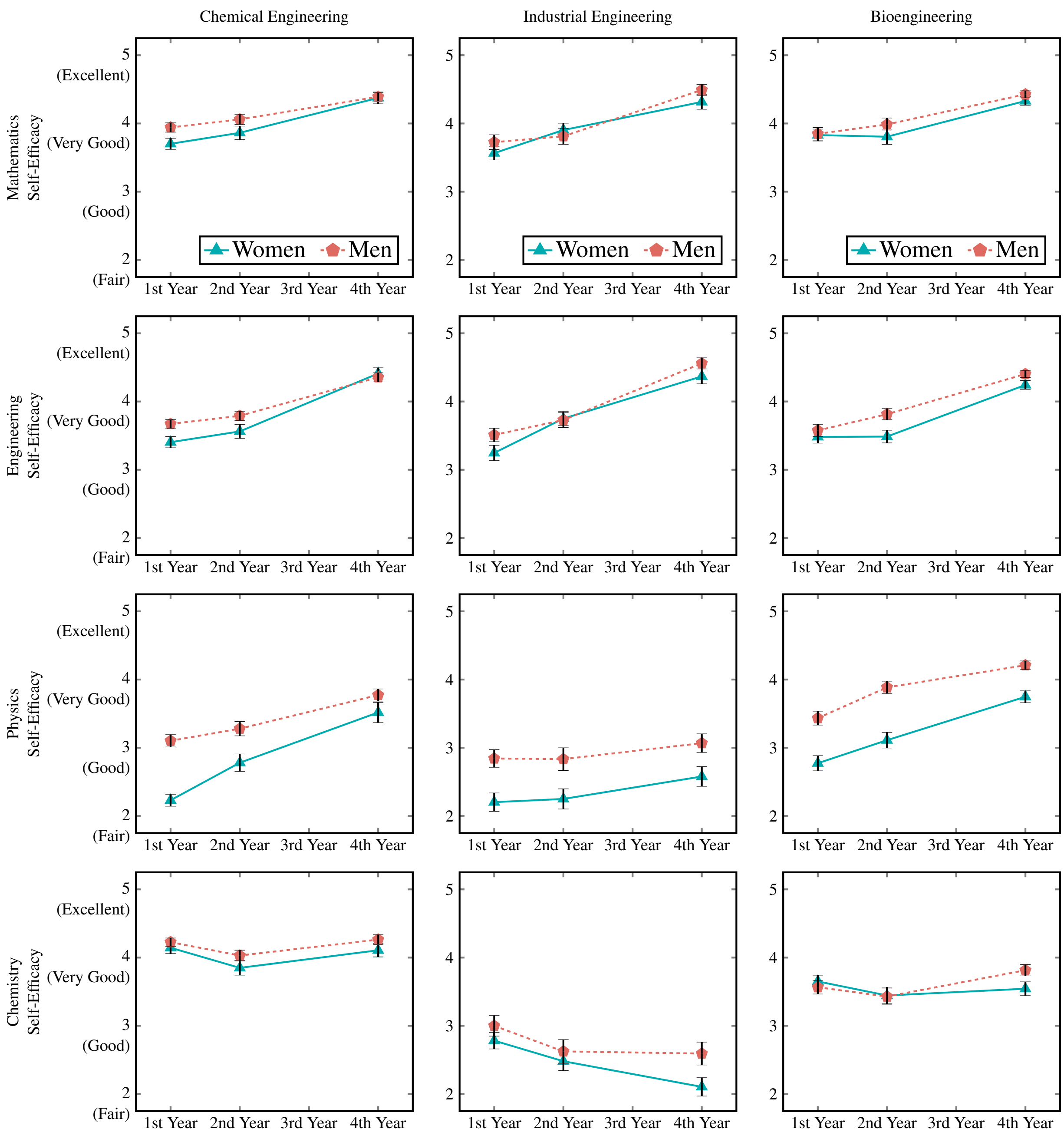
Figure 49: Distribution of responses to the physics self-efficacy prompts the surveys taken by students in their (a) 1st year, (b) 2nd year, and (c) 3rd year. The fraction and standard error of men and women, respectively, who answered each of the five options are plotted.

B.2 Self-Efficacy Over Time by Major



(Caption on next page.)

Figure 50: (Previous page.) As in Fig. 13, the mean self-efficacy scores of engineering students at the end of their first, second, and fourth years in each of the foundational subjects in engineering are plotted along with their standard error. These results are now plotted separately for students in each of the six majors (the three majors with the lowest percentage of women in this figure and the three with the highest percentages of women in Fig. 48). Self-efficacy was measured on a Likert scale from 1 to 5. Each column contains the graphs for the different majors while each row contains the graphs for self-efficacy in the different foundational subjects.



(Caption on next page.)

Figure 51: (Previous page.) As in Fig. 13, the mean self-efficacy scores of engineering students at the end of their first, second, and fourth years in each of the foundational subjects in engineering are plotted along with their standard error. These results are now plotted separately for students in each of the six majors (the three majors with the lowest percentage of women in Figure 5 and the three with the highest percentages of women in this figure). Self-efficacy was measured on a Likert scale from 1 to 5. Each column contains the graphs for the different majors while each row contains the graphs for self-efficacy in the different foundational subjects.

Appendix C Supplementary Material: Gender Analysis

C.1 Number of Majors by Term

Table 23: For each term from 1 to 12, the current number of declared majors ("Current") is shown along with the number of current majors who newly declared in that term ("Added") and the number of former majors who dropped the major as of that term ("Dropped"). In square brackets next to each measure is the percentage of all unique students who declared that major. The three sub-tables show this information for three different majors: (a) biology and neuroscience, (b) computer science, and (c) engineering.

(a) **Biological Sciences**, $N_{\text{unique}} = 3132$

Term	Number of Majors [% of N_{unique}] Current	Added	Dropped
1	24 [0.8]	24 [0.8]	0 [0.0]
2	127 [4.1]	106 [3.4]	4 [0.1]
3	1720 [54.9]	1599 [51.1]	15 [0.5]
4	2507 [80.0]	878 [28.0]	108 [3.4]
5	2810 [89.7]	390 [12.5]	99 [3.2]
6	2782 [88.8]	74 [2.4]	109 [3.5]
7	2697 [86.1]	44 [1.4]	102 [3.3]
8	2578 [82.3]	13 [0.4]	62 [2.0]
9	485 [15.5]	9 [0.3]	44 [1.4]
10	295 [9.4]	2 [0.1]	20 [0.6]
11	60 [1.9]	2 [0.1]	9 [0.3]
12	18 [0.6]	0 [0.0]	11 [0.4]

(b) **Computer Science**, $N_{\text{unique}} = 624$

Term	Number of Majors [% of N_{unique}] Current	Added	Dropped
1	1 [0.2]	1 [0.2]	0 [0.0]
2	21 [3.4]	20 [3.2]	0 [0.0]
3	101 [16.2]	82 [13.1]	2 [0.3]
4	215 [34.5]	116 [18.6]	5 [0.8]
5	397 [63.6]	189 [30.3]	8 [1.3]
6	461 [73.9]	75 [12.0]	11 [1.8]
7	503 [80.6]	64 [10.3]	17 [2.7]
8	497 [79.6]	27 [4.3]	11 [1.8]
9	242 [38.8]	37 [5.9]	10 [1.6]
10	148 [23.7]	11 [1.8]	11 [1.8]
11	69 [11.1]	2 [0.3]	10 [1.6]
12	36 [5.8]	2 [0.3]	2 [0.3]

(c) **Engineering**, $N_{\text{unique}} = 3587$

Term	Number of Majors [% of N_{unique}] Current	Added	Dropped
1	3236 [90.2]	3236 [90.2]	48 [1.3]
2	2986 [83.2]	14 [0.4]	264 [7.4]
3	2976 [83.0]	161 [4.5]	171 [4.8]
4	2911 [81.2]	56 [1.6]	121 [3.4]
5	2911 [81.2]	64 [1.8]	64 [1.8]
6	2909 [81.1]	27 [0.8]	29 [0.8]
7	2890 [80.6]	22 [0.6]	40 [1.1]
8	2856 [79.6]	5 [0.1]	25 [0.7]
9	1616 [45.1]	4 [0.1]	15 [0.4]
10	640 [17.8]	1 [0.0]	15 [0.4]
11	116 [3.2]	0 [0.0]	8 [0.2]
12	51 [1.4]	0 [0.0]	7 [0.2]

Table 24: For each term from 1 to 12, the current number of declared majors ("Current") is shown along with the number of current majors who newly declared in that term ("Added") and the number of former majors who dropped the major as of that term ("Dropped"). In square brackets next to each measure is the percentage of all unique students who declared that major. The four sub-tables show this information for four different majors: (a) mathematics, (b) chemistry, (c) physics and astronomy, and (d) geology and environmental science.

(a) **Mathematics**, $N_{\text{unique}} = 453$

Term	Number of Majors [% of N_{unique}]		
	Current	Added	Dropped
1	24 [5.3]	24 [5.3]	0 [0.0]
2	109 [24.1]	86 [19.0]	3 [0.7]
3	221 [48.8]	122 [26.9]	12 [2.6]
4	317 [70.0]	117 [25.8]	23 [5.1]
5	343 [75.7]	52 [11.5]	28 [6.2]
6	353 [77.9]	27 [6.0]	19 [4.2]
7	335 [74.0]	18 [4.0]	24 [5.3]
8	317 [70.0]	4 [0.9]	16 [3.5]
9	95 [21.0]	9 [2.0]	12 [2.6]
10	69 [15.2]	2 [0.4]	4 [0.9]
11	22 [4.9]	1 [0.2]	7 [1.5]
12	7 [1.5]	0 [0.0]	6 [1.3]

(b) **Chemistry**, $N_{\text{unique}} = 557$

Term	Number of Majors [% of N_{unique}]		
	Current	Added	Dropped
1	14 [2.5]	14 [2.5]	0 [0.0]
2	102 [18.3]	88 [15.8]	1 [0.2]
3	328 [58.9]	231 [41.5]	7 [1.3]
4	426 [76.5]	128 [23.0]	34 [6.1]
5	445 [79.9]	55 [9.9]	37 [6.6]
6	430 [77.2]	18 [3.2]	34 [6.1]
7	413 [74.1]	12 [2.2]	25 [4.5]
8	393 [70.6]	2 [0.4]	14 [2.5]
9	130 [23.3]	8 [1.4]	10 [1.8]
10	83 [14.9]	1 [0.2]	8 [1.4]
11	19 [3.4]	0 [0.0]	2 [0.4]
12	11 [2.0]	1 [0.2]	2 [0.4]

(c) **Physics & Astronomy**, $N_{\text{unique}} = 187$

Term	Number of Majors [% of N_{unique}]		
	Current	Added	Dropped
1	17 [9.1]	17 [9.1]	0 [0.0]
2	75 [40.1]	59 [31.6]	1 [0.5]
3	124 [66.3]	62 [33.2]	15 [8.0]
4	151 [80.7]	34 [18.2]	8 [4.3]
5	148 [79.1]	8 [4.3]	12 [6.4]
6	144 [77.0]	4 [2.1]	8 [4.3]
7	134 [71.7]	2 [1.1]	12 [6.4]
8	128 [68.4]	1 [0.5]	6 [3.2]
9	46 [24.6]	1 [0.5]	4 [2.1]
10	38 [20.3]	0 [0.0]	2 [1.1]
11	3 [1.6]	0 [0.0]	1 [0.5]
12	1 [0.5]	0 [0.0]	2 [1.1]

(d) **Geology**, $N_{\text{unique}} = 350$

Term	Number of Majors [% of N_{unique}]		
	Current	Added	Dropped
1	7 [2.0]	7 [2.0]	0 [0.0]
2	45 [12.9]	38 [10.9]	1 [0.3]
3	173 [49.4]	132 [37.7]	6 [1.7]
4	269 [76.9]	104 [29.7]	10 [2.9]
5	301 [86.0]	45 [12.9]	13 [3.7]
6	306 [87.4]	11 [3.1]	7 [2.0]
7	303 [86.6]	8 [2.3]	7 [2.0]
8	291 [83.1]	3 [0.9]	3 [0.9]
9	82 [23.4]	2 [0.6]	5 [1.4]
10	33 [9.4]	0 [0.0]	6 [1.7]
11	8 [2.3]	0 [0.0]	2 [0.6]
12	4 [1.1]	0 [0.0]	2 [0.6]

Table 25: For each term from 1 to 12, the current number of declared majors ("Current") is shown along with the number of majors who newly declared in that term ("Added") and the number of former majors who dropped the major as of that term ("Dropped"). In square brackets next to each measure is the percentage of all unique students who declared that major (or cluster of majors as in (e)). The three sub-tables show this information for different non-STEM majors or clusters of majors: (a) psychology, (b) economics, and (c) all other non-STEM majors.

(a) **Psychology**, $N_{\text{unique}} = 1796$

Term	Number of Majors [% of N_{unique}]		
	Current	Added	Dropped
1	7 [0.4]	7 [0.4]	0 [0.0]
2	64 [3.6]	57 [3.2]	0 [0.0]
3	408 [22.7]	345 [19.2]	4 [0.2]
4	1015 [56.5]	621 [34.6]	26 [1.4]
5	1456 [81.1]	464 [25.8]	23 [1.3]
6	1603 [89.3]	174 [9.7]	30 [1.7]
7	1664 [92.7]	96 [5.3]	13 [0.7]
8	1559 [86.8]	23 [1.3]	28 [1.6]
9	265 [14.8]	18 [1.0]	14 [0.8]
10	138 [7.7]	4 [0.2]	10 [0.6]
11	32 [1.8]	0 [0.0]	6 [0.3]
12	16 [0.9]	0 [0.0]	5 [0.3]

(b) **Economics**, $N_{\text{unique}} = 1083$

Term	Number of Majors [% of N_{unique}]		
	Current	Added	Dropped
1	20 [1.8]	20 [1.8]	0 [0.0]
2	123 [11.4]	106 [9.8]	6 [0.6]
3	367 [33.9]	252 [23.3]	9 [0.8]
4	651 [60.1]	303 [28.0]	26 [2.4]
5	820 [75.7]	203 [18.7]	35 [3.2]
6	867 [80.1]	86 [7.9]	41 [3.8]
7	856 [79.0]	68 [6.3]	59 [5.4]
8	799 [73.8]	32 [3.0]	33 [3.0]
9	209 [19.3]	11 [1.0]	22 [2.0]
10	109 [10.1]	3 [0.3]	15 [1.4]
11	36 [3.3]	3 [0.3]	6 [0.6]
12	13 [1.2]	0 [0.0]	8 [0.7]

(c) **Non-STEM**, $N_{\text{unique}} = 5346$

Term	Number of Majors [% of N_{unique}]		
	Current	Added	Dropped
1	263 [4.9]	263 [4.9]	0 [0.0]
2	1047 [19.6]	816 [15.3]	41 [0.8]
3	2514 [47.0]	1511 [28.3]	52 [1.0]
4	3636 [68.0]	1211 [22.7]	115 [2.2]
5	4302 [80.5]	807 [15.1]	94 [1.8]
6	4510 [84.4]	313 [5.9]	77 [1.4]
7	4528 [84.7]	247 [4.6]	120 [2.2]
8	4290 [80.2]	134 [2.5]	115 [2.2]
9	838 [15.7]	48 [0.9]	69 [1.3]
10	469 [8.8]	16 [0.3]	33 [0.6]
11	119 [2.2]	5 [0.1]	15 [0.3]
12	60 [1.1]	3 [0.1]	13 [0.2]

C.2 Summary Counts for Each Major

Table 26: Summary counts for all students. For each major, the total number of unique students is listed along with peak concurrent majors, added majors, and dropped majors, as well as the term in which the peak occurs in brackets. For example, in biological sciences, there were 3132 individual students in the sample who had ever declared the major. 2810 of those students had a declared biological science major in term 5 (peak term for concurrent majors), which is higher than the number of majors declared in any other term. Further, 1599 of those students added the major in term 3 (peak term for adding this major) and 109 of those students dropped the major in term 6 (peak term for dropping this major).

All Students Major	Unique Majors	Peak Concurrent Majors [Term]	Peak Added Majors [Term]	Peak Dropped Majors [Term]
Biological Sciences	3132	2810 [5]	1599 [3]	109 [6]
Computer Science	624	503 [7]	189 [5]	17 [7]
Engineering	3587	3236 [1]	3236 [1]	264 [2]
Mathematics	453	353 [6]	122 [3]	28 [5]
Chemistry	557	445 [5]	231 [3]	37 [5]
Physics and Astronomy	187	151 [4]	62 [3]	15 [3]
Geology	350	306 [6]	132 [3]	13 [5]
Economics	1083	867 [6]	303 [4]	59 [7]
Psychology	1796	1664 [7]	621 [4]	30 [6]
Other Non-STEM	5346	4528 [7]	1511 [3]	120 [7]

Table 27: Summary counts for men. For each major, the total number of unique male students is listed along with peak concurrent majors, added majors, and dropped majors, as well as the term in which the peak occurs in brackets.

Men Only Major	Unique Majors	Peak Concurrent Majors [Term]	Peak Added Majors [Term]	Peak Dropped Majors [Term]
Biological Sciences	1554	1381 [5]	829 [3]	58 [4]
Computer Science	513	413 [7]	156 [5]	14 [7]
Engineering	2683	2414 [1]	2414 [1]	190 [2]
Mathematics	295	229 [6]	72 [4]	18 [7]
Chemistry	315	253 [6]	132 [3]	19 [7]
Physics and Astronomy	151	124 [4]	54 [3]	11 [7]
Geology	153	133 [6]	60 [3]	6 [5]
Economics	737	575 [6]	203 [4]	40 [7]
Psychology	495	449 [7]	142 [4]	14 [8]
Other Non-STEM	2328	1906 [7]	615 [3]	68 [7]

Table 28: Summary counts for women. For each major, the total number of unique female students is listed along with peak concurrent majors, added majors, and dropped majors, as well as the term in which the peak occurs in brackets.

Women Only Major	Unique Majors	Peak Concurrent Majors [Term]	Peak Added Majors [Term]	Peak Dropped Majors [Term]
Biological Sciences	1578	1429 [5]	770 [3]	67 [6]
Computer Science	111	90 [7]	33 [5]	3 [7]
Engineering	904	822 [1]	822 [1]	74 [2]
Mathematics	158	124 [6]	51 [3]	13 [5]
Chemistry	242	194 [5]	99 [3]	21 [4]
Physics and Astronomy	36	28 [6]	13 [2]	4 [3]
Geology	197	173 [6]	72 [3]	7 [5]
Economics	346	292 [6]	100 [4]	19 [7]
Psychology	1301	1215 [7]	479 [4]	20 [6]
Other Non-STEM	3018	2622 [7]	896 [3]	60 [4]

C.3 Degrees Earned by Students who Dropped a Major

Table 29: Trajectory of all students who dropped a major. For each major, the total number of students in the dataset who dropped that major (N_{drop}) is listed along with the percentage of N_{drop} who ultimately earned a degree in each major or earned no degree. For example, there were 583 students who ever dropped their major in biological sciences. Of those 458 students, 2.9% went on to earn a degree in biological sciences (i.e., they later declared that major again after dropping it at an earlier point). Similarly, 3.4% of them earned a degree in computer science, 8.2% in engineering, 2.2% in mathematics, 3.4% in chemistry, and so forth. Finally, 27.1% of those 458 students that dropped a biological sciences major ultimately did not earn a degree from the university.

All Students		% of N_{drop} in a Given Major That Subsequently Earned Degree in Each Major										
Major	N_{drop}	Bio	CS	Engr	Math	Chem	Phys	Geo	Econ	Psych	Non-STEM	No Degree
Bio	583	2.9	3.4	8.2	2.2	3.4	0.2	1.7	12.2	2.7	45.8	27.1
CS	87	6.9	1.1	2.3	4.6	0.0	1.1	0.0	2.3	6.9	26.4	52.9
Engr	807	5.8	8.2	0.9	2.9	2.6	0.7	2.1	2.5	5.8	29.7	46.3
Math	154	9.1	4.5	5.8	4.5	4.5	1.3	1.3	7.1	28.6	26.0	26.0
Chem	174	16.7	4.0	6.3	5.2	6.9	1.1	5.2	6.3	2.3	43.7	23.0
Phys	71	2.8	4.2	11.3	16.9	4.2	1.4	1.4	2.8	4.2	23.9	36.6
Geo	62	11.3	0.0	0.0	0.0	1.6	0.0	1.6	6.5	1.6	35.5	53.2
Econ	260	4.6	3.5	3.8	10.8	1.5	0.0	2.3	5.0	0.8	47.7	30.4
Psych	159	3.8	3.1	0.6	0.0	0.0	0.6	0.0	10.1	1.9	37.1	50.3
Non-STEM	745	8.6	4.3	2.6	2.1	1.5	0.9	2.0	11.7	5.2	9.0	59.6

Table 30: Trajectory of all men who dropped a major. For each major, the total number of male students in the dataset who dropped that major (N_{drop}) is listed along with the percentage of N_{drop} who ultimately earned a degree in each major or earned no degree.

Men Only		% of N_{drop} in a Given Major That Subsequently Earned Degree in Each Major										
Major	N_{drop}	Bio	CS	Engr	Math	Chem	Phys	Geo	Econ	Psych	Non-STEM	No Degree
Bio	283	3.2	6.0	13.8	1.8	5.3	0.4	2.1	7.1	2.8	35.3	31.8
CS	72	8.3	1.4	2.8	4.2	0.0	1.4	0.0	0.0	4.2	27.8	55.6
Engr	611	3.8	9.7	0.5	2.8	2.0	0.8	2.3	1.3	6.1	28.6	49.3
Math	103	7.8	4.9	6.8	4.9	5.8	1.9	1.0	3.9	27.2	22.3	31.1
Chem	91	16.5	7.7	11.0	6.6	11.0	1.1	4.4	5.5	3.3	34.1	23.1
Phys	60	3.3	5.0	11.7	16.7	5.0	1.7	1.7	1.7	5.0	21.7	36.7
Geo	32	9.4	0.0	0.0	0.0	0.0	0.0	3.1	6.2	3.1	37.5	53.1
Econ	199	4.0	4.5	4.0	10.1	1.0	0.0	2.5	3.5	1.0	45.7	32.7
Psych	64	0.0	7.8	0.0	0.0	0.0	1.6	0.0	12.5	3.1	37.5	45.3
Non-STEM	408	7.8	5.9	2.9	2.7	1.5	1.2	1.0	5.6	6.4	7.8	63.2

Table 31: Trajectory of all women who dropped a major. For each major, the total number of female students in the dataset who dropped that major (N_{drop}) is listed along with the percentage of N_{drop} who ultimately earned a degree in each major or earned no degree.

Women Only		% of N_{drop} in a Given Major That Subsequently Earned Degree in Each Major										
Major	N_{drop}	Bio	CS	Engr	Math	Chem	Phys	Geo	Econ	Psych	Non-STEM	No Degree
Bio	300	2.7	1.0	3.0	2.7	1.7	0.0	1.3	17.0	2.7	55.7	22.7
CS	15	0.0	0.0	0.0	6.7	0.0	0.0	0.0	13.3	20.0	20.0	40.0
Engr	196	12.2	3.6	2.0	3.1	4.6	0.5	1.5	6.1	5.1	33.2	37.2
Math	51	11.8	3.9	3.9	3.9	2.0	0.0	2.0	13.7	31.4	33.3	15.7
Chem	83	16.9	0.0	1.2	3.6	2.4	1.2	6.0	7.2	1.2	54.2	22.9
Phys	11	0.0	0.0	9.1	18.2	0.0	0.0	0.0	9.1	0.0	36.4	36.4
Geo	30	13.3	0.0	0.0	0.0	3.3	0.0	0.0	6.7	0.0	33.3	53.3
Econ	61	6.6	0.0	3.3	13.1	3.3	0.0	1.6	9.8	0.0	54.1	23.0
Psych	95	6.3	0.0	1.1	0.0	0.0	0.0	0.0	8.4	1.1	36.8	53.7
Non-STEM	337	9.5	2.4	2.1	1.5	1.5	0.6	3.3	19.0	3.9	10.4	55.2

Appendix D Supplementary Material: Race/Ethnicity Analysis

D.1 Number of Majors by Term

Table 32: For each term from 1 to 12, the current number of declared majors ("Current") is shown along with the number of current majors who newly declared in that term ("Added") and the number of former majors who dropped the major as of that term ("Dropped"). In square brackets next to each measure is the percentage of all unique students who declared that major. The three sub-tables show this information for three different majors: (a) biology and neuroscience, (b) computer science, and (c) engineering.

(a) **Biological Sciences**, $N_{\text{unique}} = 3077$

Term	Number of Majors [% of N_{unique}] Current	Added	Dropped
1	24 [0.8]	24 [0.8]	0 [0.0]
2	126 [4.1]	105 [3.4]	4 [0.1]
3	1689 [54.9]	1569 [51.0]	15 [0.5]
4	2458 [79.9]	860 [27.9]	107 [3.5]
5	2760 [89.7]	386 [12.5]	96 [3.1]
6	2732 [88.8]	73 [2.4]	108 [3.5]
7	2648 [86.1]	43 [1.4]	100 [3.2]
8	2534 [82.4]	13 [0.4]	60 [1.9]
9	477 [15.5]	9 [0.3]	43 [1.4]
10	289 [9.4]	2 [0.1]	19 [0.6]
11	58 [1.9]	2 [0.1]	9 [0.3]
12	17 [0.6]	0 [0.0]	11 [0.4]

(b) **Computer Science**, $N_{\text{unique}} = 606$

Term	Number of Majors [% of N_{unique}] Current	Added	Dropped
1	1 [0.2]	1 [0.2]	0 [0.0]
2	21 [3.5]	20 [3.3]	0 [0.0]
3	98 [16.2]	79 [13.0]	2 [0.3]
4	209 [34.5]	113 [18.6]	5 [0.8]
5	389 [64.2]	185 [30.5]	6 [1.0]
6	452 [74.6]	74 [12.2]	11 [1.8]
7	490 [80.9]	60 [9.9]	17 [2.8]
8	485 [80.0]	26 [4.3]	10 [1.7]
9	234 [38.6]	36 [5.9]	10 [1.7]
10	144 [23.8]	11 [1.8]	10 [1.7]
11	67 [11.1]	1 [0.2]	9 [1.5]
12	34 [5.6]	2 [0.3]	2 [0.3]

(c) **Engineering**, $N_{\text{unique}} = 3515$

Term	Number of Majors [% of N_{unique}] Current	Added	Dropped
1	3171 [90.2]	3171 [90.2]	47 [1.3]
2	2924 [83.2]	14 [0.4]	261 [7.4]
3	2914 [82.9]	159 [4.5]	169 [4.8]
4	2848 [81.0]	54 [1.5]	120 [3.4]
5	2848 [81.0]	62 [1.8]	62 [1.8]
6	2847 [81.0]	27 [0.8]	28 [0.8]
7	2829 [80.5]	21 [0.6]	38 [1.1]
8	2795 [79.5]	5 [0.1]	25 [0.7]
9	1592 [45.3]	4 [0.1]	15 [0.4]
10	630 [17.9]	1 [0.0]	15 [0.4]
11	113 [3.2]	0 [0.0]	7 [0.2]
12	49 [1.4]	0 [0.0]	7 [0.2]

Table 33: For each term from 1 to 12, the current number of declared majors ("Current") is shown along with the number of current majors who newly declared in that term ("Added") and the number of former majors who dropped the major as of that term ("Dropped"). In square brackets next to each measure is the percentage of all unique students who declared that major. The four sub-tables show this information for four different majors: (a) mathematics, (b) chemistry, (c) physics and astronomy, and (d) geology and environmental science.

(d) **Mathematics**, $N_{\text{unique}} = 444$

Term	Number of Majors [% of N_{unique}]		
	Current	Added	Dropped
1	22 [5.0]	22 [5.0]	0 [0.0]
2	104 [23.4]	83 [18.7]	3 [0.7]
3	215 [48.4]	121 [27.3]	12 [2.7]
4	310 [69.8]	116 [26.1]	23 [5.2]
5	335 [75.5]	51 [11.5]	28 [6.3]
6	344 [77.5]	26 [5.9]	19 [4.3]
7	326 [73.4]	18 [4.1]	24 [5.4]
8	308 [69.4]	4 [0.9]	16 [3.6]
9	93 [20.9]	9 [2.0]	12 [2.7]
10	67 [15.1]	2 [0.5]	4 [0.9]
11	22 [5.0]	1 [0.2]	7 [1.6]
12	7 [1.6]	0 [0.0]	6 [1.4]

(a) **Chemistry**, $N_{\text{unique}} = 548$

Term	Number of Majors [% of N_{unique}]		
	Current	Added	Dropped
1	13 [2.4]	13 [2.4]	0 [0.0]
2	100 [18.2]	87 [15.9]	1 [0.2]
3	324 [59.1]	229 [41.8]	7 [1.3]
4	420 [76.6]	125 [22.8]	33 [6.0]
5	439 [80.1]	54 [9.9]	36 [6.6]
6	424 [77.4]	17 [3.1]	33 [6.0]
7	407 [74.3]	12 [2.2]	25 [4.6]
8	387 [70.6]	2 [0.4]	14 [2.6]
9	127 [23.2]	8 [1.5]	10 [1.8]
10	82 [15.0]	1 [0.2]	8 [1.5]
11	18 [3.3]	0 [0.0]	2 [0.4]
12	10 [1.8]	1 [0.2]	2 [0.4]

(e) **Physics & Astronomy**, $N_{\text{unique}} = 183$

Term	Number of Majors [% of N_{unique}]		
	Current	Added	Dropped
1	17 [9.3]	17 [9.3]	0 [0.0]
2	73 [39.9]	57 [31.1]	1 [0.5]
3	122 [66.7]	62 [33.9]	15 [8.2]
4	148 [80.9]	33 [18.0]	7 [3.8]
5	144 [78.7]	7 [3.8]	12 [6.6]
6	140 [76.5]	4 [2.2]	8 [4.4]
7	131 [71.6]	2 [1.1]	11 [6.0]
8	125 [68.3]	1 [0.5]	6 [3.3]
9	46 [25.1]	1 [0.5]	4 [2.2]
10	38 [20.8]	0 [0.0]	2 [1.1]
11	3 [1.6]	0 [0.0]	1 [0.5]
12	1 [0.5]	0 [0.0]	2 [1.1]

(d) **Geology**, $N_{\text{unique}} = 342$

Term	Number of Majors [% of N_{unique}]		
	Current	Added	Dropped
1	7 [2.0]	7 [2.0]	0 [0.0]
2	45 [13.2]	38 [11.1]	1 [0.3]
3	168 [49.1]	127 [37.1]	6 [1.8]
4	262 [76.6]	102 [29.8]	10 [2.9]
5	294 [86.0]	44 [12.9]	12 [3.5]
6	300 [87.7]	11 [3.2]	6 [1.8]
7	297 [86.8]	8 [2.3]	7 [2.0]
8	285 [83.3]	3 [0.9]	3 [0.9]
9	80 [23.4]	2 [0.6]	5 [1.5]
10	33 [9.6]	0 [0.0]	6 [1.8]
11	8 [2.3]	0 [0.0]	2 [0.6]
12	4 [1.2]	0 [0.0]	2 [0.6]

Table 34: For each term from 1 to 12, the current number of declared majors ("Current") is shown along with the number of majors who newly declared in that term ("Added") and the number of former majors who dropped the major as of that term ("Dropped"). In square brackets next to each measure is the percentage of all unique students who declared that major (or cluster of majors as in (e)). The three sub-tables show this information for different non-STEM majors or clusters of majors: (a) psychology, (b) economics, and (c) all other non-STEM majors.

(c) **Psychology**, $N_{\text{unique}} = 1779$

Term	Number of Majors [% of N_{unique}] Current	Added	Dropped
1	7 [0.4]	7 [0.4]	0 [0.0]
2	63 [3.5]	56 [3.1]	0 [0.0]
3	404 [22.7]	342 [19.2]	4 [0.2]
4	1005 [56.5]	615 [34.6]	26 [1.5]
5	1446 [81.3]	464 [26.1]	23 [1.3]
6	1590 [89.4]	171 [9.6]	30 [1.7]
7	1650 [92.7]	95 [5.3]	13 [0.7]
8	1543 [86.7]	21 [1.2]	28 [1.6]
9	262 [14.7]	18 [1.0]	14 [0.8]
10	134 [7.5]	3 [0.2]	10 [0.6]
11	30 [1.7]	0 [0.0]	6 [0.3]
12	15 [0.8]	0 [0.0]	4 [0.2]

(b) **Economics**, $N_{\text{unique}} = 1048$

Term	Number of Majors [% of N_{unique}] Current	Added	Dropped
1	18 [1.7]	18 [1.7]	0 [0.0]
2	117 [11.2]	102 [9.7]	6 [0.6]
3	350 [33.4]	241 [23.0]	9 [0.9]
4	629 [60.0]	297 [28.3]	25 [2.4]
5	792 [75.6]	196 [18.7]	34 [3.2]
6	836 [79.8]	83 [7.9]	41 [3.9]
7	827 [78.9]	67 [6.4]	57 [5.4]
8	771 [73.6]	31 [3.0]	32 [3.1]
9	205 [19.6]	11 [1.0]	21 [2.0]
10	106 [10.1]	3 [0.3]	15 [1.4]
11	35 [3.3]	3 [0.3]	6 [0.6]
12	12 [1.1]	0 [0.0]	8 [0.8]

(e) **Non-STEM**, $N_{\text{unique}} = 5246$

Term	Number of Majors [% of N_{unique}] Current	Added	Dropped
1	258 [4.9]	258 [4.9]	0 [0.0]
2	1021 [19.5]	795 [15.2]	40 [0.8]
3	2468 [47.0]	1491 [28.4]	52 [1.0]
4	3563 [67.9]	1184 [22.6]	115 [2.2]
5	4218 [80.4]	793 [15.1]	91 [1.7]
6	4424 [84.3]	309 [5.9]	75 [1.4]
7	4445 [84.7]	241 [4.6]	117 [2.2]
8	4214 [80.3]	132 [2.5]	112 [2.1]
9	816 [15.6]	48 [0.9]	68 [1.3]
10	457 [8.7]	14 [0.3]	33 [0.6]
11	117 [2.2]	4 [0.1]	14 [0.3]
12	59 [1.1]	3 [0.1]	12 [0.2]

D.2 Summary Counts for Each Major

Table 35: Summary counts for all students. For each major, the total number of unique students is listed along with peak concurrent majors, added majors, and dropped majors, as well as the term in which the peak occurs in brackets. For example, in biological sciences (including neuroscience), there were 3077 individual students in the sample who had ever declared the major. 2760 of those students declared biological science majors in term 5 (peak term for concurrent majors), which is higher than the number of majors declared in any other term. Further, 1569 of those students added the major in term 3 (peak term for adding this major) and 108 of those students dropped the major in term 6 (peak term for dropping this major).

All Students Major	Unique Majors	Max Simultaneous Majors [Term]	Max Added Majors [Term]	Max Dropped Majors [Term]
Biological Sciences	3077	2760 [5]	1569 [3]	108 [6]
Computer Science	606	490 [7]	185 [5]	17 [7]
Engineering	3515	3171 [1]	3171 [1]	261 [2]
Mathematics	444	344 [6]	121 [3]	28 [5]
Chemistry	548	439 [5]	229 [3]	36 [5]
Physics and Astronomy	183	148 [4]	62 [3]	15 [3]
Geology	342	300 [6]	127 [3]	12 [5]
Economics	1048	836 [6]	297 [4]	57 [7]
Psychology	1779	1650 [7]	615 [4]	30 [6]
Other Non-STEM	5246	4445 [7]	1491 [3]	117 [7]

Table 36: Summary counts for Asian students. For each major, the total number of unique Asian students is listed along with peak concurrent majors, added majors, and dropped majors, as well as the term in which the peak occurs in brackets.

Asian Students Major	Unique Majors	Max Simultaneous Majors [Term]	Max Added Majors [Term]	Max Dropped Majors [Term]
Biological Sciences	651	597 [5]	358 [3]	26 [6]
Computer Science	93	74 [8]	20 [5]	4 [7]
Engineering	296	256 [3]	255 [1]	16 [2]
Mathematics	63	47 [6]	20 [3]	6 [4]
Chemistry	77	60 [6]	26 [3]	8 [4]
Physics and Astronomy	11	9 [4]	5 [4]	2 [5]
Geology	15	14 [7]	5 [3]	1 [9]
Economics	206	172 [6]	63 [3]	14 [7]
Psychology	209	191 [7]	70 [4]	6 [6]
Other Non-STEM	494	401 [7]	124 [3]	16 [7]

Table 37: Summary counts for URM students. For each major, the total number of unique URM students is listed along with peak concurrent majors, added majors, and dropped majors, as well as the term in which the peak occurs in brackets.

URM Students Major	Unique Majors	Max Simultaneous Majors [Term]	Max Added Majors [Term]	Max Dropped Majors [Term]
Biological Sciences	288	248 [5]	121 [3]	15 [5]
Computer Science	72	63 [8]	30 [5]	4 [11]
Engineering	348	313 [1]	313 [1]	28 [2]
Mathematics	30	25 [5]	8 [3]	2 [7]
Chemistry	41	28 [5]	14 [3]	4 [7]
Physics and Astronomy	17	15 [5]	11 [3]	3 [7]
Geology	24	23 [7]	8 [3]	1 [12]
Economics	112	82 [7]	35 [4]	8 [8]
Psychology	269	244 [7]	86 [5]	9 [6]
Other Non-STEM	642	531 [7]	159 [3]	20 [8]

Table 38: Summary counts for White students. For each major, the total number of unique White students is listed along with peak concurrent majors, added majors, and dropped majors, as well as the term in which the peak occurs in brackets.

White Students Major	Unique Majors	Max Simultaneous Majors [Term]	Max Added Majors [Term]	Max Dropped Majors [Term]
Biological Sciences	2138	1915 [5]	1090 [3]	78 [4]
Computer Science	441	363 [7]	135 [5]	12 [7]
Engineering	2871	2603 [1]	2603 [1]	217 [2]
Mathematics	351	273 [6]	96 [4]	26 [5]
Chemistry	430	355 [5]	189 [3]	30 [5]
Physics and Astronomy	155	126 [4]	51 [2]	13 [3]
Geology	303	264 [6]	114 [3]	11 [5]
Economics	730	585 [6]	214 [4]	37 [7]
Psychology	1301	1215 [7]	464 [4]	22 [8]
Other Non-STEM	4110	3513 [7]	1208 [3]	85 [4]

D.3 Degrees Earned by Students who Dropped a Major

Table 39: Trajectory of all students who dropped a major. For each major, the total number of students in the dataset who dropped that major (N_{drop}) is listed along with the percentage of N_{drop} who ultimately earned a degree in each major or earned no degree. For example, there were 572 students who ever dropped their major in biological sciences. Of those 572 students, 2.8% went on to earn a degree in biological sciences (i.e., they later declared that major again after dropping it at an earlier point). Similarly, 3.3% of them earned a degree in computer science, 8.4% in engineering, 2.3% in mathematics, 3.5% in chemistry, and so forth. Finally, 27.3% of those 572 students that dropped a biological sciences major ultimately did not earn a degree from the university.

All Students		% of N_{drop} in a Given Major That Subsequently Earned Degree in Each Major										
Major	N_{drop}	Bio	CS	Engr	Math	Chem	Phys	Geo	Econ	Other Psych	Non-STEM	No Degree
Bio	572	2.8	3.3	8.4	2.3	3.5	0.2	1.6	12.2	2.8	45.5	27.3
CS	82	6.1	1.2	2.4	4.9	0.0	1.2	0.0	2.4	4.9	25.6	53.7
Engr	794	5.9	8.2	0.9	2.9	2.6	0.8	2.1	2.4	5.8	30.0	46.1
Math	154	9.1	4.5	5.8	4.5	4.5	1.3	1.3	7.1	28.6	26.0	26.0
Chem	171	16.4	4.1	5.8	4.7	6.4	1.2	5.3	6.4	2.3	44.4	23.4
Phys	69	2.9	4.3	11.6	17.4	2.9	1.4	1.4	2.9	4.3	24.6	36.2
Geo	60	11.7	0.0	0.0	0.0	1.7	0.0	1.7	6.7	1.7	35.0	53.3
Econ	254	4.7	3.5	3.5	11.0	1.2	0.0	2.4	5.1	0.8	47.6	30.3
Psych	158	3.8	3.2	0.6	0.0	0.0	0.6	0.0	10.1	1.9	37.3	50.0
Non-STEM	730	8.6	4.2	2.6	2.2	1.5	1.0	2.1	11.9	4.8	8.5	59.9

Table 40: Trajectory of all Asian students who dropped a major. For each major, the total number of Asian students in the dataset who dropped that major (N_{drop}) is listed along with the percentage of N_{drop} who ultimately earned a degree in each major or earned no degree.

Asian Students		% of N_{drop} in a Given Major That Subsequently Earned Degree in Each Major										
Major	N_{drop}	Bio	CS	Engr	Math	Chem	Phys	Geo	Econ	Other Psych	Non-STEM	No Degree
Bio	121	6.6	5.8	8.3	2.5	1.7	0.0	0.0	9.9	6.6	42.1	28.1
CS	13	7.7	0.0	0.0	0.0	0.0	0.0	0.0	7.7	7.7	23.1	61.5
Engr	67	9.0	13.4	0.0	1.5	3.0	1.5	0.0	3.0	10.4	23.9	47.8
Math	20	10.0	5.0	5.0	0.0	0.0	0.0	0.0	10.0	40.0	15.0	35.0
Chem	25	24.0	4.0	8.0	4.0	12.0	0.0	0.0	8.0	0.0	32.0	32.0
Phys	5	0.0	0.0	20.0	20.0	0.0	0.0	0.0	0.0	0.0	40.0	20.0
Geo	2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	50.0	0.0	0.0	50.0
Econ	48	8.3	4.2	2.1	6.2	4.2	0.0	0.0	12.5	0.0	31.2	39.6
Psych	14	7.1	0.0	0.0	0.0	0.0	0.0	0.0	7.1	0.0	50.0	35.7
Non-STEM	65	16.9	3.1	3.1	1.5	3.1	0.0	0.0	15.4	4.6	4.6	53.8

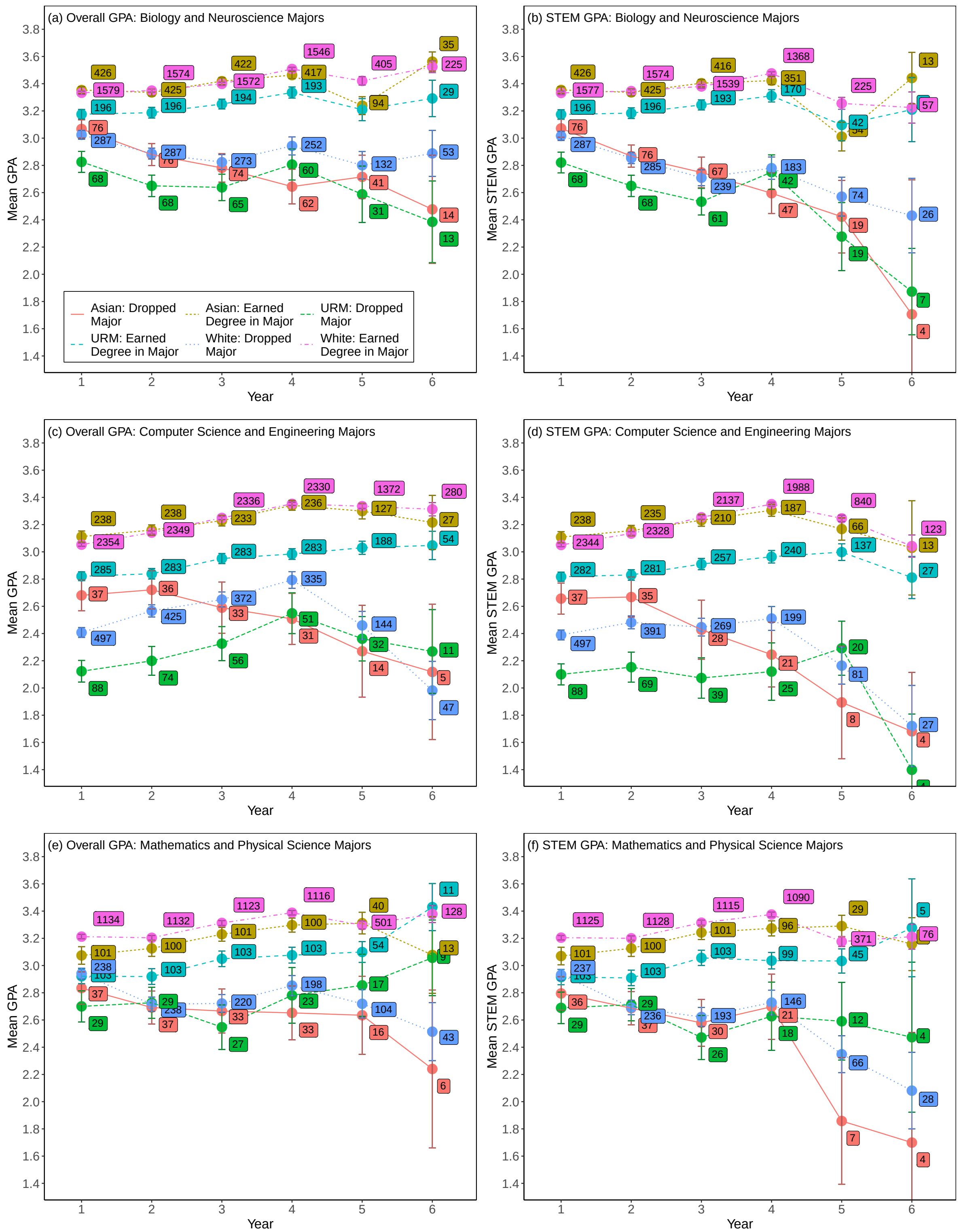
Table 41: Trajectory of all URM students who dropped a major. For each major, the total number of URM students in the dataset who dropped that major (N_{drop}) is listed along with the percentage of N_{drop} who ultimately earned a degree in each major or earned no degree.

URM Students		% of N_{drop} in a Given Major That Subsequently Earned Degree in Each Major										
Major	N_{drop}	Bio	CS	Engr	Math	Chem	Phys	Geo	Econ	Other Psych	Non-STEM	No Degree
Bio	78	2.6	0.0	1.3	2.6	2.6	0.0	2.6	9.0	1.3	57.7	25.6
CS	7	14.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	28.6	57.1
Engr	98	3.1	7.1	0.0	4.1	1.0	0.0	1.0	2.0	1.0	29.6	54.1
Math	10	0.0	0.0	20.0	10.0	0.0	0.0	0.0	10.0	30.0	40.0	30.0
Chem	19	5.3	10.5	10.5	10.5	5.3	0.0	0.0	5.3	5.3	52.6	15.8
Phys	11	9.1	0.0	9.1	18.2	0.0	9.1	0.0	9.1	9.1	27.3	18.2
Geo	5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	40.0	60.0
Econ	41	0.0	4.9	7.3	9.8	0.0	0.0	2.4	7.3	0.0	46.3	31.7
Psych	29	3.4	6.9	0.0	0.0	0.0	0.0	0.0	6.9	0.0	31.0	62.1
Non-STEM	127	8.7	2.4	0.8	0.0	0.0	0.8	0.0	7.9	2.4	8.7	72.4

Table 42: Trajectory of all White students who dropped a major. For each major, the total number of White students in the dataset who dropped that major (N_{drop}) is listed along with the percentage of N_{drop} who ultimately earned a degree in each major or earned no degree.

White Students		% of N_{drop} in a Given Major That Subsequently Earned Degree in Each Major										
Major	N_{drop}	Bio	CS	Engr	Math	Chem	Phys	Geo	Econ	Other Psych	Non-STEM	No Degree
Bio	373	1.6	3.2	9.9	2.1	4.3	0.3	1.9	13.7	1.9	44.0	27.3
CS	62	4.8	1.6	3.2	6.5	0.0	1.6	0.0	1.6	4.8	25.8	51.6
Engr	629	6.0	7.8	1.1	2.9	2.9	0.8	2.5	2.4	6.0	30.7	44.7
Math	124	9.7	4.8	4.8	4.8	5.6	1.6	1.6	6.5	26.6	26.6	24.2
Chem	127	16.5	3.1	4.7	3.9	5.5	1.6	7.1	6.3	2.4	45.7	22.8
Phys	53	1.9	5.7	11.3	17.0	3.8	0.0	1.9	1.9	3.8	22.6	41.5
Geo	53	13.2	0.0	0.0	0.0	1.9	0.0	1.9	5.7	1.9	35.8	52.8
Econ	165	4.8	3.0	3.0	12.7	0.6	0.0	3.0	2.4	1.2	52.7	27.3
Psych	115	3.5	2.6	0.9	0.0	0.0	0.9	0.0	11.3	2.6	37.4	48.7
Non-STEM	538	7.6	4.8	3.0	2.8	1.7	1.1	2.8	12.5	5.4	8.9	57.6

D.4 GPA by Major



(Caption on next page.)

Figure 52: GPA and STEM GPA over time for STEM majors by racial/ethnic group. Each GPA is calculated yearly, not cumulatively. Majors are divided into three groupings: (a) and (b) biological sciences and neuroscience; (c) and (d) computer science and engineering; and (e) and (f) mathematics and physical science (chemistry, physics and astronomy, and geology and environmental science). GPA in all courses – (a), (c), and (e) – and in only STEM courses – (b), (d), and (f) – are calculated separately for four categories of students that declared at least one of the majors in each group: Asian, URM and White students that ultimately earned a degree in that group of majors and those that dropped from that group of majors. For each group, the mean GPA is plotted along with its standard error, with the sample size listed above each point and guides to the eye connecting the points.

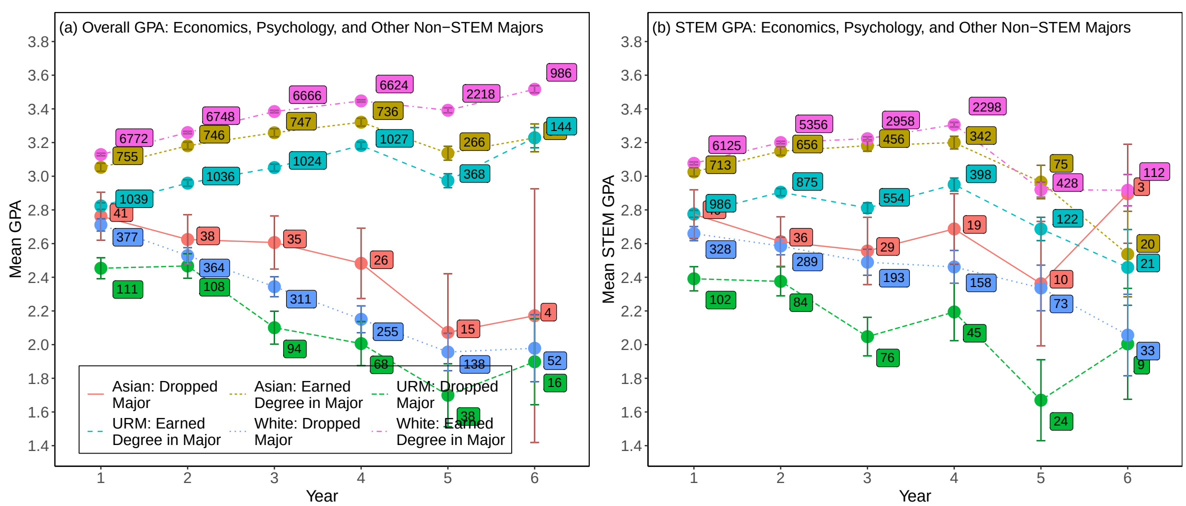


Figure 53: GPA and STEM GPA over time for non-STEM majors by racial/ethnic group. Each GPA is calculated yearly, not cumulatively. GPAs in (a) all courses and (b) in only STEM courses are calculated separately for four categories of students that declared at least one of the majors in each group: Asian, URM and White students that ultimately earned a degree in that group of majors and those that dropped from that group of majors. For each group, the mean GPA is plotted along with its standard error, with the sample size listed above each point and guides to the eye connecting the points.

Bibliography

[1] N. Abramzon, P. Benson, E. Bertschinger, S. Blessing, G. L. Cochran, A. Cox, B. Cunningham, J. Galbraith-Frew, J. Johnson, L. Kerby, E. Lalanne, C. O'Donnell, S. Petty, S. Sampath, S. Seestrom, C. Singh, C. Spencer, K. S. Woodle, and S. Yennello. Women in physics in the United States: Recruitment and retention. *AIP Conference Proceedings*, 1697(1):060045, 2015.

[2] B. A. 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, Jun 2012.

[3] P. Ashby and F. M. Mensah. Critical chemistry education in a private, suburban high school. *Research in Science Education*, 50(1):303–332, 2020.

[4] A. W. Astin et al. *What Matters in College*, volume 9. Jossey-Bass, 1993.

[5] T. J. Atherton, R. S. Barthelemy, W. Deconinck, M. L. Falk, S. Garmon, E. Long, M. Plisch, E. H. Simmons, and K. Reeves. *LGBT climate in physics: Building an inclusive community*. American Physical Society, College Park, MD, 2016.

[6] R. S. Baker and P. S. Inventado. Educational data mining and learning analytics. In *Learning Analytics*, pages 61–75. Springer, 2014.

[7] S. F. Bancroft. Toward a critical theory of science, technology, engineering, and mathematics doctoral persistence: Critical capital theory. *Science Education*, 102(6):1319–1335, 2018.

[8] A. Bandura. Social cognitive theory of self-regulation. *Organizational Behavior and Human Decision Processes*, 50(2):248–287, 1991.

[9] A. Bandura. Self-efficacy. In R. J. Corsini, editor, *Encyclopedia of Psychology*, volume 3, pages 368–369. Wiley, 1994.

[10] A. Bandura. *Self-efficacy: The Exercise of Control*. Macmillan, 1997.

[11] A. Bandura. Social cognitive theory of personality. In L. A. Pervin and O. P. John, editors, *Handbook of Personality*, volume 2, pages 154–196. Oxford University Press, 1999.

[12] A. Bandura. Social cognitive theory: An agentic perspective. *Annual Review of Psychology*, 52(1):1–26, 2001.

[13] A. Bandura. The evolution of social cognitive theory. In K. G. Smith and M. A. Hitt, editors, *Great Minds in Management*, pages 9–35. Oxford University Press, 2005.

[14] M. Bang and D. Medin. Cultural processes in science education: Supporting the navigation of multiple epistemologies. *Science Education*, 94(6):1008–1026, 2010.

[15] V. Basile and E. Lopez. And still I see no changes: Enduring views of students of color in science and mathematics education policy reports. *Science Education*, 99(3):519–548, 2015.

[16] M. Bassok and K. J. Holyoak. Interdomain transfer between isomorphic topics in algebra and physics. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15(1):153, 1989.

[17] S. Bates, R. Donnelly, C. MacPhee, D. Sands, M. Birch, and N. R. Walet. Gender differences in conceptual understanding of newtonian mechanics: a UK cross-institution comparison. *European Journal of Physics*, 34(2):421–434, Feb 2013.

[18] S. L. Beilock, R. J. Rydell, and A. R. McConnell. Stereotype threat and working memory: mechanisms, alleviation, and spillover. *Journal of Experimental Psychology: General*, 136(2):256, 2007.

[19] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, 57(1):289–300, 1995.

[20] Z. Berry, I. Turgeon, and F. Yonekura. *Obojobo Next*, Jun 2019.

[21] M. Besterfield-Sacre, C. J. Atman, and L. J. Shuman. Characteristics of freshman engineering students: Models for determining student attrition in engineering. *Journal of Engineering Education*, 86(2):139–149, 1997.

[22] M. Besterfield-Sacre, M. Moreno, L. J. Shuman, and C. J. Atman. Gender and ethnicity differences in freshmen engineering student attitudes: A cross-institutional study. *Journal of Engineering Education*, 90(4):477–489, 2001.

[23] L. Betancur, E. Votruba-Drzal, and C. Schunn. Socioeconomic gaps in science achievement. *International Journal of STEM Education*, 5(1):38, 2018.

[24] N. E. Betz and G. Hackett. Applications of self-efficacy theory to understanding career choice behavior. *Journal of Social and Clinical Psychology*, 4(3):279–289, 1986.

[25] L. Bian. *The roots of gender gaps: Investigating the development of gender stereotypes about intelligence*. PhD thesis, University of Illinois at Urbana-Champaign, 2017.

[26] L. Bian, S.-J. Leslie, and A. Cimpian. Gender stereotypes about intellectual ability emerge early and influence children's interests. *Science*, 355(6323):389–391, 2017.

[27] L. Bian, S.-J. Leslie, and A. Cimpian. Evidence of bias against girls and women in contexts that emphasize intellectual ability. *American Psychologist*, 73(9):1139, 2018.

[28] L. Bian, S.-J. Leslie, M. C. Murphy, and A. Cimpian. Messages about brilliance undermine women's interest in educational and professional opportunities. *Journal of Experimental Social Psychology*, 76:404–420, 2018.

[29] J. A. Bianchini. Expanding underrepresented minority participation: America's science and technology talent at the crossroads. *Science Education*, 97(1):163–166, 2013.

[30] J. A. Bianchini, C. C. Johnston, S. Y. Oram, and L. M. Cavazos. Learning to teach science in contemporary and equitable ways: The successes and struggles of first-year science teachers. *Science Education*, 87(3):419–443, 2003.

[31] J. A. Bianchini, D. J. Whitney, T. D. Breton, and B. A. Hilton-Brown. Toward inclusive science education: University scientists' views of students, instructional practices, and the nature of science. *Science Education*, 86(1):42–78, 2002.

[32] J. Bilak and C. Singh. Improving students' conceptual understanding of conductors and insulators. *AIP Conference Proceedings*, 951(1):49–52, 2007.

[33] K. R. Binning, N. Kaufmann, E. McGreevy, O. Fotuhi, S. Chen, E. Marshman, Z. Y. Kalender, L. Limeri, L. Betancur, and C. Singh. Changing social norms to foster the benefits of collaboration in diverse workgroups. unpublished, N.D.

[34] J. A. Birt, M. Khajeloo, C. C. Rega-Brodsky, M. A. Siegel, T. S. Hancock, K. Cummings, and P. D. Nguyen. Fostering agency to overcome barriers in college science teaching: Going against the grain to enact reform-based ideas. *Science Education*, 103(4):770–798, 2019.

[35] R. Bjorkquist, A. M. Bogdan, N. L. Campbell, M. Chessey, G. L. Cochran, B. Cunningham, J. N. Esquivel, L. Gladstone, N. M. Gosnell, S. Guruswamy, K. M. Hallinen, C. Harris, A. Johnson, J. L. Johnson, C. Jones, R. A. Jorgenson, L. McCullough, M. D. McNeese, T. D. Presley, N. Quist, A. Richardson, S. Seidel, and C. Singh. Women in physics in the United States: Reaching toward equity and inclusion. *AIP Conference Proceedings*, 2109(1):050040, 2019.

[36] M. Boekaerts and M. Niemivirta. Self-regulated learning: Finding a balance between learning goals and ego-protective goals. In M. Boekaerts, P. R. Pintrich, and M. Zeidner, editors, *Handbook of Self-Regulation*, pages 417 – 450. Academic Press, San Diego, 2000.

[37] M. Borrego and J. Bernhard. The emergence of engineering education research as an internationally connected field of inquiry. *Journal of Engineering Education*, 100(1):14–47, 2011.

[38] M. Borrego and C. Henderson. Increasing the use of evidence-based teaching in STEM higher education: A comparison of eight change strategies. *Journal of Engineering Education*, 103(2):220–252, 2014.

[39] M. Borrego, R. A. Streveler, R. L. Miller, and K. A. Smith. A new paradigm for a new field: Communicating representations of engineering education research. *Journal of Engineering Education*, 97(2):147–162, 2008.

[40] U. Bossér and M. Lindahl. Students' positioning in the classroom: A study of teacher-student interactions in a socioscientific issue context. *Research in Science Education*, 49(2):371–390, 2019.

[41] M. C. Bottia, E. Stearns, R. A. Mickelson, and S. Moller. Boosting the numbers of STEM majors? the role of high schools with a STEM program. *Science Education*, 102(1):85–107, 2018.

[42] S. G. Brainard and L. Carlin. A six-year longitudinal study of undergraduate women in engineering and science. *Journal of Engineering Education*, 87(4):369–375, 1998.

[43] J. D. Bransford and D. L. Schwartz. Rethinking transfer: A simple proposal with multiple implications. *Review of Research in Education*, 24:61–100, 1999.

[44] C. E. Brawner, M. M. Camacho, S. M. Lord, R. A. Long, and M. W. Ohland. Women in industrial engineering: Stereotypes, persistence, and perspectives. *Journal of Engineering Education*, 101(2):288–318, 2012.

[45] C. E. Brawner, S. M. Lord, R. A. Layton, M. W. Ohland, and R. A. Long. Factors affecting women's persistence in chemical engineering. *International Journal of Engineering Education*, 31(6A):1431–1447, 2015.

[46] S. L. Britner and F. Pajares. Sources of science self-efficacy beliefs of middle school students. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching*, 43(5):485–499, 2006.

[47] H. S. Broudy. Types of knowledge and purposes of education. In R. C. Anderson, R. J. Spiro, and W. E. Montague, editors, *Schooling and the Acquisition of Knowledge*, pages 1–17. Routledge, 1977.

[48] M. W. Browne, R. Cudeck, et al. Alternative ways of assessing model fit. *Sage Focus Editions*, 154:136–136, 1993.

[49] M. A. Cannady, D. Moore, E. Votruba-Drzal, E. Greenwald, R. Stites, and C. D. Schunn. How personal, behavioral, and environmental factors predict working in STEMM vs non-STEMM middle-skill careers. *International Journal of STEM Education*, 4(1):22, 2017.

[50] H. B. Carlone and A. Johnson. Understanding the science experiences of successful women of color: Science identity as an analytic lens. *Journal of Research in Science Teaching*, 44(8):1187–1218, 2007.

[51] M. Carlson and S. A. Mulaik. Trait ratings from descriptions of behavior as mediated by components of meaning. *Multivariate Behavioral Research*, 28(1):111–159, 1993.

[52] N. K. Cauthen and S. Fass. Measuring income and poverty in the United States. Technical report, National Center for Children in Poverty, 2007.

[53] A. M. Cavallo, W. H. Potter, and M. Rozman. Gender differences in learning constructs, shifts in learning constructs, and their relationship to course achievement in a structured inquiry, yearlong college physics course for life science majors. *School Science and Mathematics*, 104(6):288–300, 2004.

[54] L. Charleston, R. P. Adserias, N. M. Lang, and J. F. Jackson. Intersectionality and STEM: The role of race and gender in the academic pursuits of African American women in STEM. *Journal of Progressive Policy & Practice*, 2(3):273–293, 2014.

[55] Z. Chen, C. Chudzicki, D. Palumbo, G. Alexandron, Y.-J. Choi, Q. Zhou, and D. E. Pritchard. Researching for better instructional methods using AB experiments in MOOCs: Results and challenges. *Research and Practice in Technology Enhanced Learning*, 11(1), Apr 2016.

[56] Z. Chen, G. Garrido, Z. Berry, I. Turgeon, and F. Yonekura. Designing online learning modules to conduct pre- and post-testing at high frequency. In *Physics Education Research Conference 2017*, PER Conference, pages 84–87, Cincinnati, OH, July 26-27 2017. Physics Education Research Topical Group and the American Association of Physics Teachers.

[57] Z. Chen and G. Gladding. How to make a good animation: A grounded cognition model of how visual representation design affects the construction of abstract physics knowledge. *Physical Review ST Physics Education Research*, 10:010111, Apr 2014.

[58] Z. Chen, S. Lee, and G. Garrido. Re-designing the structure of online courses to empower educational data mining. In *Proceedings of the 11th International Conference on Educational Data Mining, EDM 2018, Buffalo, NY, USA, July 15-18, 2018*, 2018.

[59] Z. Chen, K. M. Whitcomb, M. W. Guthrie, and C. Singh. Evaluating the effectiveness of two methods to improve students' problem solving performance after studying an online tutorial. In *Physics Education Research Conference 2019*, PER Conference, Provo, UT, July 24-25 2019. Physics Education Research Topical Group and the American Association of Physics Teachers.

[60] Z. Chen, K. M. Whitcomb, and C. Singh. Measuring the effectiveness of online problem-solving tutorials by multi-level knowledge transfer. In A. Traxler, Y. Cao, and S. Wolf, editors, *Physics Education Research Conference 2018*, PER Conference, Washington, DC, 2018. Physics Education Research Topical Group and the American Association of Physics Teachers.

[61] Z. Chen, M. Xu, G. Garrido, and M. W. Guthrie. Relationship between students' online learning behavior and course performance: What contextual information matters? *Physical Review Physics Education Research*, 16:010138, Jun 2020.

[62] S. Cheryan and V. C. Plaut. Explaining underrepresentation: A theory of precluded interest. *Sex Roles*, 63(7-8):475–488, 2010.

[63] S. Cheryan, S. A. Ziegler, A. K. Montoya, and L. Jiang. Why are some STEM fields more gender balanced than others? *Psychological Bulletin*, 143(1):1, 2017.

[64] S. Cho, K. W. Crenshaw, and L. McCall. Toward a field of intersectionality studies: Theory, applications, and praxis. *Signs: Journal of Women in Culture and Society*, 38(4):785–810, 2013.

[65] J. Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, 2 edition, 1988.

[66] V. P. Coletta, J. A. Phillips, A. Savinainen, and J. J. Steinert. Comment on 'the effects of students' reasoning abilities on conceptual understanding and problem-solving skills in introductory mechanics'. *European Journal of Physics*, 29(5):L25–L27, Jul 2008.

[67] M. P. Coupland, A. P. Gardner, and G. Carmody Jones. Mathematics for engineering education: What students say. In *Proceedings of the 31st MERGA Conference*. Mathematics Education Research Group of Australasia, 2008.

[68] K. Crenshaw. Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stan. L. Rev.*, 43:1241, 1990.

[69] K. Crenshaw, N. Gotanda, G. Peller, and K. Thomas. *Critical Race Theory: The Key Writings that Formed the Movement*. The New Press, 1995.

[70] K. P. Cross. On college teaching. *Journal of Engineering Education*, 82(1):9–14, 1993.

[71] S. G. Daley. Towards inclusion of all learners through science teacher education. *Science Education*, 103(5):1306–1308, 2019.

[72] M. H. Dancy and C. Henderson. Pedagogical practices and instructional change of physics faculty. *American Journal of Physics*, 78(10):1056–1063, 2010.

[73] A. T. Danielsson. Exploring woman university physics students 'doing gender' and 'doing physics'. *Gender and Education*, 24(1):25–39, 2012.

[74] A. T. Danielsson and C. Linder. Learning in physics by doing laboratory work: Towards a new conceptual framework. *Gender and Education*, 21(2):129–144, 2009.

[75] S. B. Day and R. L. Goldstone. The import of knowledge export: Connecting findings and theories of transfer of learning. *Educational Psychologist*, 47(3):153–176, 2012.

[76] F. F. M. De Mul, C. M. i Batlle, I. De Bruijn, and K. Rinzema. How to encourage university students to solve physics problems requiring mathematical skills: The 'adventurous problem solving' approach. *European Journal of Physics*, 25(1):51–61, Oct 2003.

[77] M.-N. Delisle, F. Guay, C. Senécal, and S. Larose. Predicting stereotype endorsement and academic motivation in women in science programs: A longitudinal model. *Learning and Individual Differences*, 19(4):468–475, 2009.

[78] L. A. DeSilva, A. Pullen, and J. E. Hasbun. Enhancing student performance in introductory physics in topics related to electricity and magnetism through the use of voluntary workshops. *European Journal of Physics*, 39(3):035702, Mar 2018.

[79] D. K. Detterman and R. J. Sternberg. *Transfer on trial: Intelligence, cognition, and instruction.* Ablex Publishing, 1993.

[80] S. DeVore, E. Marshman, and C. Singh. Challenge of engaging all students via self-paced interactive electronic learning tutorials for introductory physics. *Physical Review Physics Education Research*, 13(1):010127, 2017.

[81] S. L. Dika and M. M. D'Amico. Early experiences and integration in the persistence of first-generation college students in STEM and non-STEM majors. *Journal of Research in Science Teaching*, 53(3):368–383, 2016.

[82] R. Dou, Z. Hazari, K. Dabney, G. Sonnert, and P. Sadler. Early informal STEM experiences and STEM identity: The importance of talking science. *Science Education*, 103(3):623–637, 2019.

[83] D. Doucette, R. Clark, and C. 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.

[84] D. Doucette and C. Singh. Why are there so few women in physics? reflections on the experiences of two women. *The Physics Teacher*, 58(5):297–300, 2020.

[85] J. Dunlosky, K. A. Rawson, E. J. Marsh, M. J. Nathan, and D. T. Willingham. Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, 14(1):4–58, 2013. PMID: 26173288.

[86] J. S. Eccles. Understanding women's educational and occupational choices: Applying the Eccles et al. model of achievement-related choices. *Psychology of Women Quarterly*, 18(4):585–609, 1994.

[87] J. S. Eccles, J. E. Jacobs, and R. D. Harold. Gender role stereotypes, expectancy effects, and parents' socialization of gender differences. *Journal of Social Issues*, 46(2):183–201, 1990.

[88] J. Eccles (Parsons), T. Adler, and J. Meece. Sex differences in achievement: A test of alternative theories. *Journal of Personality and Social Psychology*, 46(1):26–43, 1984.

[89] S. L. Eddy and K. A. Hogan. Getting under the hood: How and for whom does increasing course structure work? *CBE Life Sciences Education*, 13(3):453–468, 2014.

[90] C. Efthimiou, D. Maronde, T. McGreevy, E. del Barco, and S. McCole. Implementing elements of The Physics Suite at a large metropolitan research university. *Physics Education*, 46(4):421–429, Jun 2011.

[91] R. A. Engle. Framing interactions to foster generative learning: A situative explanation of transfer in a community of learners classroom. *The Journal of the Learning Sciences*, 15(4):451–498, 2006.

[92] O. Eris, D. Chachra, H. L. Chen, S. Sheppard, L. Ludlow, C. Rosca, T. Bailey, and G. Toye. Outcomes of a longitudinal administration of the persistence in engineering survey. *Journal of Engineering Education*, 99(4):371–395, 2010.

[93] M. Estrada, A. Eroy-Reveles, and J. Matsui. The influence of affirming kindness and community on broadening participation in STEM career pathways. *Social Issues and Policy Review*, 12(1):258–297, 2018.

[94] R. M. Felder, G. N. Felder, and E. J. Dietz. A longitudinal study of engineering student performance and retention. V. Comparisons with traditionally-taught students. *Journal of Engineering Education*, 87(4):469–480, 1998.

[95] R. M. Felder, G. N. Felder, M. Mauney, C. E. Hamrin Jr., and E. J. Dietz. A longitudinal study of engineering student performance and retention. III. Gender differences in student performance and attitudes. *Journal of Engineering Education*, 84(2):151–163, 1995.

[96] J. Flegg, D. Mallet, and M. Lupton. Students' perceptions of the relevance of mathematics in engineering. *International Journal of Mathematical Education in Science and Technology*, 43:717–732, 09 2012.

[97] J. M. Fraser, A. L. Timan, K. Miller, J. E. Dowd, L. Tucker, and E. Mazur. Teaching and physics education research: Bridging the gap. *Reports on Progress in Physics*, 77(3):032401, Mar 2014.

[98] D. Freedman, R. Pisani, and R. Purves. *Statistics*. W. W. Norton & Co., 4 edition, 2007.

[99] C. M. Ganley, C. E. George, J. R. Cimpian, and M. 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.

[100] V. Gibson, L. Jardine-Wright, and E. 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, May 2015.

[101] M. L. Gick and K. J. Holyoak. Analogical problem solving. *Cognitive Psychology*, 12(3):306–355, 1980.

[102] A. J. Gonsalves. Exploring how gender figures the identity trajectories of two doctoral students in observational astrophysics. *Physical Review Physics Education Research*, 14:010146, Jun 2018.

[103] C. Good, J. Aronson, and M. 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.

[104] C. Good, A. Rattan, and C. 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.

[105] M. Good, A. Maries, and C. 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*, 15:020129, Sep 2019.

[106] M. Gopalan and A. A. Nelson. Understanding the racial discipline gap in schools. *AERA Open*, 5(2):2332858419844613, 2019.

[107] A. M. Green, B. R. Brand, and G. E. Glasson. Applying actor–network theory to identify factors contributing to nonpersistence of African American students in STEM majors. *Science Education*, 103(2):241–263, 2019.

[108] B. Gregorcic and J. Haglund. Conceptual blending as an interpretive lens for student engagement with technology: Exploring celestial motion on an interactive whiteboard. *Research in Science Education*,, 2018.

[109] M. W. Guthrie, T. Zhang, and Z. Chen. A tale of two guessing strategies: Interpreting the time students spend solving problems through online log data. Submitted to Physics Education Research Conference 2020, 2020.

[110] R. Gutiérrez. Framing equity: Helping students "play the game" and "change the game". *Teaching for Excellence and Equity in Mathematics*, 1(1):4–8, 2009.

[111] L. Hamilton. Statistics with stata 12, 2013.

[112] K. A. Harper, L. Abrams, and J. P. Ruffley. A longitudinal study of the impact of a first-year honors engineering program. In *2014 ASEE Annual Conference & Exposition*, Indianapolis, Indiana, June 2014. ASEE Conferences.

[113] K. A. Harper, G. R. Baker, and D. M. Grzybowski. First steps in strengthening the connections between mathematics and engineering. In *2013 ASEE Annual Conference & Exposition*, Atlanta, Georgia, June 2013. ASEE Conferences.

[114] Z. Hazari, G. Potvin, R. M. Lock, F. Lung, G. Sonnert, and P. M. Sadler. Factors that affect the physical science career interest of female students: Testing five common hypotheses. *Physical Review ST Physics Education Research*, 9:020115, Oct 2013.

[115] Z. Hazari, P. M. Sadler, and G. 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.

[116] Z. Hazari, G. Sonnert, P. M. Sadler, and M.-C. Shanahan. Connecting high school physics experiences, outcome expectations, physics identity, and physics career choice: A gender study. *Journal of Research in Science Teaching*, 47(8):978–1003, 2010.

[117] Z. Hazari, R. H. Tai, and P. M. Sadler. Gender differences in introductory university physics performance: The influence of high school physics preparation and affective factors. *Science Education*, 91(6):847–876, 2007.

[118] C. Henderson, M. Dancy, and M. Niewiadomska-Bugaj. Use of research-based instructional strategies in introductory physics: Where do faculty leave the innovation-decision process? *Physical Review ST Physics Education Research*, 8:020104, Jul 2012.

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

[120] C. Henderson, J. P. Mestre, and L. L. Slakey. Cognitive science research can improve undergraduate STEM instruction: What are the barriers? *Policy Insights from the Behavioral and Brain Sciences*, 2(1):51–60, 2015.

[121] R. Henderson, G. Stewart, J. Stewart, L. Michaluk, and A. Traxler. Exploring the gender gap in the conceptual survey of electricity and magnetism. *Physical Review Physics Education Research*, 13:020114, Sep 2017.

[122] E. Hennessey, J. Cole, P. Shastri, J. Esquivel, C. Singh, R. Johnson, and S. Ghose. Workshop report: Intersecting identities—gender and intersectionality in physics. *AIP Conference Proceedings*, 2109(1):040001, 2019.

[123] D. Hestenes, M. Wells, and G. Swackhamer. Force concept inventory. *The Physics Teacher*, 30(3):141–158, 1992.

[124] N. M. Hewitt and E. Seymour. A long, discouraging climb. *ASEE Prism*, 1(6):24–28, 1992.

[125] C. Hill, C. Corbett, and A. St. Rose. *Why So Few? Women in Science, Technology, Engineering, and Mathematics.* American Association of University Women, 2010.

[126] A. Hilts, R. Part, and M. L. Bernacki. The roles of social influences on student competence, relatedness, achievement, and retention in STEM. *Science Education*, 102(4):744–770, 2018.

[127] D. Ho, K. Imai, G. King, and E. Stuart. MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, 42(8):1–28, 2011.

[128] 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.

[129] P. M. Huang and S. G. Brainard. Identifying determinants of academic self confidence among science, math, engineering, and technology students. *Journal of Women and Minorities in Science and Engineering*, 7(4):315, 2001.

[130] S. Huang and N. Fang. Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models. *Computers & Education*, 61(1):133–145, 2013.

[131] B. E. Hughes. Coming out in STEM: Factors affecting retention of sexual minority STEM students. *Science Advances*, 4(3):eaao6373, 2018.

[132] M. M. Hurley. Reviewing integrated science and mathematics: The search for evidence and definitions from new perspectives. *School Science and Mathematics*, 101(5):259–268, 2001.

[133] Indiana University Center for Postsecondary Research. The Carnegie classification of institutions of higher education. , Indiana University Center for Postsecondary Research, Bloomington, IN, 2018.

[134] Y. Jiang, M. Ekono, and C. Skinner. Basic facts about low-income children. Technical report, National Center for Children in Poverty, 2015.

[135] A. Johansson. Negotiating intelligence, nerdiness, and status in physics master's studies. *Research in Science Education*,, 2018.

[136] A. Johnson. The mathematics of sex: How biology and society conspire to limit talented women and girls. *Science Education*, 96:960–962, 2012.

[137] A. Johnson, M. Ong, L. T. Ko, J. Smith, and A. 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.

[138] A. Johri and B. M. Olds. Situated engineering learning: Bridging engineering education research and the learning sciences. *Journal of Engineering Education*, 100(1):151–185, 2011.

[139] Z. Y. Kalender, E. Marshman, T. J. Nokes-Malach, C. Schunn, and C. Singh. Motivational characteristics of underrepresented ethnic and racial minority students in introductory physics courses. In *Physics Education Research Conference 2017*, PER Conference, pages 204–207, Cincinnati, OH, July 26-27 2017.

[140] Z. Y. Kalender, E. Marshman, C. Schunn, T. J. Nokes-Malach, and C. Singh. Large gender differences in physics self-efficacy at equal performance levels: A warning sign?

In *Physics Education Research Conference 2018*, PER Conference, Washington, DC, August 1-2 2018.

[141] Z. Y. Kalender, E. Marshman, C. Schunn, T. J. Nokes-Malach, and C. Singh. Investigating the role of prior preparation and self-efficacy on female and male students' introductory physics course achievements. In *Physics Education Research Conference 2019*, PER Conference, Provo, UT, July 24-25 2019.

[142] Z. Y. Kalender, E. Marshman, C. D. Schunn, T. J. Nokes-Malach, and C. Singh. Gendered patterns in the construction of physics identity from motivational factors. *Physical Review Physics Education Research*, 15:020119, Aug 2019.

[143] Z. Y. Kalender, E. Marshman, C. D. Schunn, T. J. Nokes-Malach, and C. Singh. Why female science, technology, engineering, and mathematics majors do not identify with physics: They do not think others see them that way. *Physical Review Physics Education Research*, 15:020148, Dec 2019.

[144] Z. Y. Kalender, E. Marshman, C. D. Schunn, T. J. Nokes-Malach, and C. Singh. Beliefs about competence: The story of self-efficacy, gender, and physics. In A. Murrell, J. Petrie-Wyman, and A. Soudi, editors, *Diversity Across Disciplines: Research on People, Policy, Process and Paradigm*, pages 3–18. Information Age Publishing, 2020.

[145] Z. Y. Kalender, E. Marshman, C. D. Schunn, T. J. Nokes-Malach, and C. Singh. Damage caused by women's lower self-efficacy on physics learning. *Physical Review Physics Education Research*, 16:010118, Apr 2020.

[146] M. Kapur. Productive failure in mathematical problem solving. *Instructional Science*, 38(6):523–550, Apr 2009.

[147] N. I. Karim, A. Maries, and C. Singh. Do evidence-based active-engagement courses reduce the gender gap in introductory physics? *European Journal of Physics*, 39(2):025701, Jan 2018.

[148] D. Karnam, K. K. Mashood, and A. Sule. Do student difficulties with vectors emerge partly from the limitations of static textbook media? *European Journal of Physics*, 41(3):035703, Apr 2020.

[149] A. V. Katrevich and M. S. Aruguete. Recognizing challenges and predicting success in first-generation university students. *Journal of STEM Education: Innovations and Research*, 18(2):40–44, 2017.

[150] D. Kellner. Toward a critical theory of education. *Democracy & Nature*, 9(1):51–64, 2003.

[151] B. King. Does postsecondary persistence in STEM vary by gender? *AERA Open*, 2(4):2332858416669709, 2016.

[152] R. B. Kline. *Principles and Practice of Structural Equation Modeling*. Guilford Press, 3 edition, 2011.

[153] E. C. Kokkelenberg, M. Dillon, and S. M. Christy. The effects of class size on student grades at a public university. *Economics of Education Review*, 27(2):221–233, April 2008.

[154] E. Kuley, S. Maw, and T. Fonstad. Understanding barriers to student success: What students have to say. In *2016 Proceedings of the CEEA*. Canadian Engineering Education Association, 2016.

[155] E. A. Kuley, S. Maw, and T. Fonstad. Engineering student retention and attrition literature review. In *2015 Proceedings of the CEEA*. Canadian Engineering Education Association, 2015.

[156] P. C. Lam, T. Srivatsan, D. Doverspike, J. Vesalo, and P. R. Mawasha. A ten year assessment of the pre-engineering program for under-represented, low income and/or first generation college students at the University of Akron. *Journal of STEM Education: Innovations and Research*, 6(3):14–20, 2005.

[157] R. W. Lent, H.-B. Sheu, D. Singley, J. A. Schmidt, L. C. Schmidt, and C. S. Gloster. Longitudinal relations of self-efficacy to outcome expectations, interests, and major choice goals in engineering students. *Journal of Vocational Behavior*, 73(2):328–335, 2008.

[158] S.-J. Leslie, A. Cimpian, M. Meyer, and E. Freeland. Expectations of brilliance underlie gender distributions across academic disciplines. *Science*, 347(6219):262–265, 2015.

[159] O. Levrini, A. D. Ambrosis, S. Hemmer, A. Laherto, M. Malgieri, O. Pantano, and G. Tasquier. Understanding first-year students' curiosity and interest about physics—lessons learned from the HOPE project. *European Journal of Physics*, 38(2):025701, Dec 2016.

[160] J. L. Lewis, H. Menzies, E. I. Nájera, and R. N. Page. Rethinking trends in minority participation in the sciences. *Science Education*, 93(6):961–977, 2009.

[161] E. Lichtenberger and C. George-Jackson. Predicting high school students' interest in majoring in a STEM field: Insight into high school students' postsecondary plans. *Journal of Career and Technical Education*, 28(1):19–38, 2013.

[162] S.-Y. Lin and C. Singh. Using an isomorphic problem pair to learn introductory physics: Transferring from a two-step problem to a three-step problem. *Physical Review ST Physics Education Research*, 9:020114, Oct 2013.

[163] T. Litzinger, L. R. Lattuca, R. Hadgraft, and W. Newstetter. Engineering education and the development of expertise. *Journal of Engineering Education*, 100(1):123–150, 2011.

[164] R. M. Lock, Z. Hazari, and G. Potvin. Impact of out-of-class science and engineering activities on physics identity and career intentions. *Physical Review Physics Education Research*, 15(2):020137, 2019.

[165] S. Lord, R. Layton, M. Ohland, C. Brawner, and R. Long. A multi-institution study of student demographics and outcomes in chemical engineering. *Chemical Engineering Education*, 48(4):231–238, 2014.

[166] S. M. Lord, M. M. Camacho, R. A. Layton, R. A. Long, M. W. Ohland, and M. H. Wasburn. Who's persisting in engineering? A comparative analysis of female and male Asian, black, Hispanic, Native American, and white students. *Journal of Women and Minorities in Science and Engineering*, 15(2):167–190, 2009.

[167] S. M. Lord, R. A. Layton, and M. W. Ohland. Multi-institution study of student demographics and outcomes in electrical and computer engineering in the USA. *IEEE Transactions on Education*, 58(3):141–150, 2015.

[168] M. Lorenzo, C. H. Crouch, and E. Mazur. Reducing the gender gap in the physics classroom. *American Journal of Physics*, 74(2):118–122, 2006.

[169] A. Madsen, S. B. McKagan, and E. C. Sayre. Gender gap on concept inventories in physics: What is consistent, what is inconsistent, and what factors influence the gap? *Physical Review ST Physics Education Research*, 9:020121, Nov 2013.

[170] J. B. Main, K. J. Mumford, and M. W. Ohland. Examining the influence of engineering students' course grades on major choice and major switching behavior. *International Journal of Engineering Education*, 31(6):1468–1475, 2015.

[171] A. V. Maltese and C. S. Cooper. STEM pathways: Do men and women differ in why they enter and exit? *AERA Open*, 3(3):2332858417727276, 2017.

[172] A. V. Maltese and R. H. Tai. Pipeline persistence: Examining the association of educational experiences with earned degrees in STEM among US students. *Science Education*, 95(5):877–907, 2011.

[173] G. C. Marchand and G. Taasoobshirazi. Stereotype threat and women's performance in physics. *International Journal of Science Education*, 35(18):3050–3061, 2013.

[174] A. Maries, N. I. Karim, and C. Singh. Is agreeing with a gender stereotype correlated with the performance of female students in introductory physics? *Physical Review Physics Education Research*, 14:020119, Nov 2018.

[175] A. Maries and C. Singh. Stereotype threat? Effects of inquiring about test takers' gender on conceptual test performance in physics. *AIP Conference Proceedings*, 1697(1):120008, 2015.

[176] R. M. Marra, K. A. Rodgers, D. Shen, and B. Bogue. Women engineering students and self-efficacy: A multi-year, multi-institution study of women engineering student self-efficacy. *Journal of Engineering Education*, 98(1):27–38, 2009.

[177] E. Marshman, Z. Y. Kalender, C. Schunn, T. Nokes-Malach, and C. Singh. A longitudinal analysis of students' motivational characteristics in introductory physics courses: Gender differences. *Canadian Journal of Physics*, 96(4):391–405, 2017.

[178] E. M. Marshman, Z. Y. Kalender, T. Nokes-Malach, C. Schunn, and C. Singh. Female students with A's have similar physics self-efficacy as male students with C's in introductory courses: A cause for alarm? *Physical Review Physics Education Research*, 14:020123, Dec 2018.

[179] M. Marušić, N. Erceg, and J. Sliško. Partially specified physics problems: University students' attitudes and performance. *European Journal of Physics*, 32(3):711–722, Mar 2011.

[180] J. Marx and K. Cummings. Development of a survey instrument to gauge students' problem-solving abilities. *AIP Conference Proceedings*, 1289(1):221–224, 2010.

[181] R. L. Matz, B. P. Koester, S. Fiorini, G. Grom, L. Shepard, C. G. Stangor, B. Weiner, and T. A. McKay. Patterns of gendered performance differences in large introductory courses at five research universities. *AERA Open*, 3(4):2332858417743754, 2017.

[182] A. Mayo, M. D. Sharma, and D. A. Muller. Qualitative differences between learning environments using videos in small groups and whole class discussions: A preliminary study in physics. *Research in Science Education*, 39(4):477–493, 2009.

[183] K. McCavit and N. E. B. Zellner. Persistence of physics and engineering students via peer mentoring, active learning, and intentional advising. *European Journal of Physics*, 37(6):065702, Sep 2016.

[184] L. C. McDermott, M. L. Rosenquist, and E. H. Van Zee. Student difficulties in connecting graphs and physics: Examples from kinematics. *American Journal of Physics*, 55(6):503–513, 1987.

[185] L. C. McDermott, P. S. Shaffer, and M. L. Rosenquist. *Physics by Inquiry: An Introduction to Physics and the Physical Sciences*, volume 1. John Wiley & Sons, 1995.

[186] B. Means, H. Wang, X. Wei, E. Iwatani, and V. Peters. Broadening participation in STEM college majors: Effects of attending a STEM-focused high school. *AERA Open*, 4(4):2332858418806305, 2018.

[187] B. Means, H. Wang, X. Wei, S. Lynch, V. Peters, V. Young, and C. Allen. Expanding STEM opportunities through inclusive STEM-focused high schools. *Science Education*, 101(5):681–715, 2017.

[188] H. Metcalf, D. Russell, and C. Hill. Broadening the science of broadening participation in STEM through critical mixed methodologies and intersectionality frameworks. *American Behavioral Scientist*, 62(5):580–599, 2018.

[189] K. L. Meyers, G. W. Bucks, K. A. Harper, and V. E. Goodrich. Multi-institutional evaluation of engineering discipline selection. In *2015 ASEE Annual Conference & Exposition*, Seattle, Washington, June 2015. ASEE Conferences.

[190] B. D. Mikula and A. F. Heckler. Framework and implementation for improving physics essential skills via computer-based practice: Vector math. *Physical Review Physics Education Research*, 13:010122, May 2017.

[191] D. I. Miller, A. H. Eagly, and M. C. Linn. Women's representation in science predicts national gender-science stereotypes: Evidence from 66 nations. *Journal of Educational Psychology*, 107(3):631, 2015.

[192] Y. Min, G. Zhang, R. A. Long, T. J. Anderson, and M. W. Ohland. Nonparametric survival analysis of the loss rate of undergraduate engineering students. *Journal of Engineering Education*, 100(2):349–373, 2011.

[193] J. D. Mitchell, C. Y. Simmons, and L. A. Greyerbiehl. *Intersectionality & Higher Education*. Peter Lang, 2014.

[194] F. Molin, S. Cabus, C. Haelermans, and W. Groot. Toward reducing anxiety and increasing performance in physics education: Evidence from a randomized experiment. *Research in Science Education,*, 2019.

[195] D. C. Montgomery, E. A. Peck, and G. G. Vining. *Introduction to Linear Regression Analysis*. John Wiley & Sons, 4 edition, 2012.

[196] M. B. Morin, K. M. Kecskemety, K. A. Harper, and P. A. Clingan. The inverted classroom in a first-year engineering course. *ASEE Conference Proceedings*, 23:1, 2013.

[197] T. R. Morton and E. C. Parsons. #BlackGirlMagic: The identity conceptualization of Black women in undergraduate STEM education. *Science Education*, 102(6):1363–1393, 2018.

[198] J. W. Mutegi, B. Sorge, G. A. Fore, and G. S. Gibau. A tale of two camps: A mixed methods investigation into racially disparate outcomes in a nanotechnology research experience. *Science Education*, 103(6):1456–1477, 2019.

[199] National Science Board. Science and engineering indicators 2018. NSB-2018-1, National Science Foundation, Alexandria, VA, 2018.

[200] National Student Clearinghouse Research Center. Science & engineering degree attainment: 2004 – 2014. , National Student Clearinghouse, 2015.

[201] J. Neter, M. H. Kutner, C. J. Nachtsheim, and W. Wasserman. *Applied Linear Statistical Models*. McGraw-Hill/Irwin, 5 edition, 2004.

[202] G. M. Nicholls, H. Wolfe, M. Besterfield-Sacre, and L. J. Shuman. Predicting STEM degree outcomes based on eighth grade data and standard test scores. *Journal of Engineering Education*, 99(3):209–223, 2010.

[203] G. M. Nicholls, H. Wolfe, M. Besterfield-Sacre, L. J. Shuman, and S. Larpkiattaworn. A method for identifying variables for predicting STEM enrollment. *Journal of Engineering Education*, 96(1):33–44, 2007.

[204] J. Nissen, R. Donatello, and B. Van Dusen. Missing data and bias in physics education research: A case for using multiple imputation. *Physical Review Physics Education Research*, 15:020106, Jul 2019.

[205] J. M. Nissen and J. T. Shemwell. Gender, experience, and self-efficacy in introductory physics. *Physical Review Physics Education Research*, 12(2):020105, 2016.

[206] J. M. Nissen, R. M. Talbot, A. Nasim Thompson, and B. Van Dusen. Comparison of normalized gain and Cohen's d for analyzing gains on concept inventories. *Physical Review Physics Education Research*, 14:010115, Mar 2018.

[207] T. J. Nokes-Malach, Z. Y. Kalender, E. Marshman, C. Schunn, and C. 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.

[208] T. J. Nokes-Malach, Z. Y. Kalender, E. Marshman, C. Schunn, and C. Singh. How is perception of being recognized by others as someone good at physics related to female and male students' physics identities? In *Physics Education Research Conference 2019*, PER Conference, Provo, UT, July 24-25 2019.

[209] T. J. Nokes-Malach, E. Marshman, Z. Y. Kalender, C. Schunn, and C. Singh. Investigation of male and female students' motivational characteristics throughout an introductory physics course sequence. In *Physics Education Research Conference 2017*, PER Conference, pages 276–279, Cincinnati, OH, July 26-27 2017.

[210] M. W. Ohland, C. E. Brawner, M. M. Camacho, R. A. Layton, R. A. Long, S. M. Lord, and M. H. Wasburn. Race, gender, and measures of success in engineering education. *Journal of Engineering Education*, 100(2):225–252, 2011.

[211] M. W. Ohland and R. A. Long. The multiple-institution database for investigating engineering longitudinal development: An experiential case study of data sharing and reuse. *Advances in Engineering Education*, 5(2):1–25, 2016.

[212] M. W. Ohland, S. D. Sheppard, G. Lichtenstein, O. Eris, D. Chachra, and R. A. Layton. Persistence, engagement, and migration in engineering programs. *Journal of Engineering Education*, 97(3):259–278, 2008.

[213] M. Ong, J. M. Smith, and L. T. Ko. Counterspaces for women of color in STEM higher education: Marginal and central spaces for persistence and success. *Journal of Research in Science Teaching*, 55(2):206–245, 2018.

[214] F. Pajares. Gender differences in mathematics self-efficacy beliefs. In A. M. Gallagher and J. C. Kaufman, editors, *Gender Differences in Mathematics: An Integrative Psychological Approach*, pages 294–315. Cambridge University Press, 2005.

[215] F. Pajares and G. Valiante. Grade level and gender differences in the writing self-beliefs of middle school students. *Contemporary Educational Psychology*, 24(4):390 – 405, 1999.

[216] F. M. Pajares. Self-efficacy beliefs, motivation, and achievement in writing: A review of the literature. *Reading & Writing Quarterly*, 19(2):139–158, 2003.

[217] Z. Papamitsiou and A. A. Economides. Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. *Journal of Educational Technology & Society*, 17(4):49–64, 2014.

[218] M. Parappilly, L. Schmidt, and S. D. Ritter. Ready to learn physics: a team-based learning model for first year university. *European Journal of Physics*, 36(5):055052, Aug 2015.

[219] M. Parappilly, R. J. Woodman, and S. Randhawa. Feasibility and effectiveness of different models of team-based learning approaches in STEMM-based disciplines. *Research in Science Education*,, 2019.

[220] D. Paunesku, G. M. Walton, C. Romero, E. N. Smith, D. S. Yeager, and C. S. Dweck. Mind-set interventions are a scalable treatment for academic underachievement. *Psychological Science*, 26(6):784–793, 2015.

[221] A. Pawl, A. Barrantes, C. Cardamone, S. Rayyan, and D. E. Pritchard. Development of a mechanics reasoning inventory. *AIP Conference Proceedings*, 1413(1):287–290, 2012.

[222] M. Pilotte, M. W. Ohland, S. M. Lord, R. A. Layton, and M. K. Orr. Student demographics, pathways, and outcomes in industrial engineering. *International Journal of Engineering Education*, 33(2):506–518, 2017.

[223] P. R. Pintrich. The role of goal orientation in self-regulated learning. In M. Boekaerts, P. R. Pintrich, and M. Zeidner, editors, *Handbook of Self-Regulation*, pages 451–502. Academic Press, San Diego, 2000.

[224] P. R. Pintrich. A motivational science perspective on the role of student motivation in learning and teaching contexts. *Journal of Educational Psychology*, 95(4):667, 2003.

[225] P. R. Pintrich. A conceptual framework for assessing motivation and self-regulated learning in college students. *Educational Psychology Review*, 16(4):385–407, 2004.

[226] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019.

[227] D. Reeping, D. Knight, J. Grohs, and S. Case. Visualization and analysis of student enrollment patterns in foundational engineering courses. *International Journal of Engineering Education*, 35(1A), Apr 2019.

[228] L. S. Richman, M. vanDellen, and W. Wood. How women cope: Being a numerical minority in a male-dominated profession. *Journal of Social Issues*, 67(3):492–509, 2011.

[229] E. E. Rigdon. CFI versus RMSEA: A comparison of two fit indexes for structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, 3(4):369–379, 1996.

[230] D. M. Riley. LGBT-friendly workplaces in engineering. *Leadership and Management in Engineering*, 8(1):19–23, 2008.

[231] I. Rodriguez, G. Potvin, and L. H. Kramer. How gender and reformed introductory physics impacts student success in advanced physics courses and continuation in the physics major. *Physical Review Physics Education Research*, 12(2):020118, 2016.

[232] K. Rosa and F. M. Mensah. Educational pathways of black women physicists: Stories of experiencing and overcoming obstacles in life. *Physical Review Physics Education Research*, 12(2):020113, 2016.

[233] B. H. Ross. This is like that: The use of earlier problems and the separation of similarity effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 13(4):629, 1987.

[234] Y. Rosseel. lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2):1–36, 2012.

[235] P. M. Sadler and R. H. Tai. The two high-school pillars supporting college science. *Science*, 317(5837):457–458, 2007.

[236] N. Safavian. What makes them persist? Expectancy-value beliefs and the math participation, performance, and preparedness of Hispanic youth. *AERA Open*, 5(3):2332858419869342, 2019.

[237] C. J. Sangwin and M. Harjula. Online assessment of dimensional numerical answers using STACK in science. *European Journal of Physics*, 38(3):035701, Feb 2017.

[238] A. Savinainen. FCI Leeds to Finland. *Physics Education*, 51(4):040105, May 2016.

[239] V. Sawtelle, E. Brewe, and L. H. Kramer. Exploring the relationship between self-efficacy and retention in introductory physics. *Journal of Research in Science Teaching*, 49(9):1096–1121, 2012.

[240] L. J. Sax, K. J. Lehman, R. S. Barthelemy, and G. Lim. Women in physics: A comparison to science, technology, engineering, and math education over four decades. *Physical Review Physics Education Research*, 12(2):020108, 2016.

[241] R. Sayer, E. Marshman, and C. Singh. Case study evaluating just-in-time teaching and peer instruction using clickers in a quantum mechanics course. *Physical Review Physics Education Research*, 12:020133, Oct 2016.

[242] K. Schenkel and A. Calabrese Barton. Critical science agency and power hierarchies: Restructuring power within groups to address injustice beyond them. *Science Education*, 104(3):500–529, 2020.

[243] C. Seron, S. S. Silbey, E. Cech, and B. Rubineau. Persistence is cultural: Professional socialization and the reproduction of sex segregation. *Work and Occupations*, 43(2):178–214, 2016.

[244] E. Seymour and N. M. Hewitt. *Talking About Leaving: Why Undergraduates Leave the Sciences*. Westview Press, 1997.

[245] M. J. Sheth. Grappling with racism as foundational practice of science teaching. *Science Education*, 103(1):37–60, 2019.

[246] H. A. Simon. *Information-processing theory of human problem solving*, chapter 1, pages 271–295. Lawrence Erlbaum, 1978.

[247] S. Singer and K. A. Smith. Discipline-based education research: Understanding and improving learning in undergraduate science and engineering. *Journal of Engineering Education*, 102(4):468–471, 2013.

[248] C. Singh. Interactive learning tutorials on quantum mechanics. *American Journal of Physics*, 76(4):400–405, 2008.

[249] C. Singh and D. Haileselassie. Developing problem-solving skills of students taking introductory physics via web-based tutorials. *Journal of College Science Teaching*, 39(4):42, 2010.

[250] C. Singh and A. Maries. Core graduate courses: A missed learning opportunity? *AIP Conference Proceedings*, 1513(1):382–385, 2013.

[251] J. L. Smith, K. L. Lewis, L. Hawthorne, and S. 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.

[252] D. Solorzano, M. Ceja, and T. Yosso. Critical race theory, racial microaggressions, and campus racial climate: The experiences of African American college students. *Journal of Negro Education*, 69:60–73, 2000.

[253] C. M. Steele. *Whistling Vivaldi: How stereotypes affect us and what we can do.* WW Norton & Company, 2011.

[254] C. M. Steele and J. Aronson. Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology*, 69(5):797, 1995.

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

[256] A. J. Swart. Does it matter which comes first in a curriculum for engineering students - Theory or practice? *International Journal of Electrical Engineering Education*, 47(2):189–199, 2010.

[257] J. Sweller, P. Ayres, and S. Kalyuga. *Cognitive Load Theory.* Springer New York, 2011.

[258] E. Taylor, D. Gillborn, and G. Ladson-Billings. *Foundations of Critical Race Theory in Education.* Routledge, 2009.

[259] R. K. Thornton and D. R. Sokoloff. Assessing student learning of Newton's laws: The force and motion conceptual evaluation and the evaluation of active learning laboratory and lecture curricula. *American Journal of Physics*, 66(4):338–352, 1998.

[260] S. Tobias. *They're Not Dumb, They're Different: Stalking the Second Tier*, volume 101. Research Corporation, 1990.

[261] S. Tolbert, A. Schindel, and A. J. Rodriguez. Relevance and relational responsibility in justice-oriented science education research. *Science Education*, 102(4):796–819, 2018.

[262] K. L. Tonso. *On the outskirts of engineering: Learning identity, gender, and power via engineering practice.* Brill Sense, Leiden, The Netherlands, 2007.

[263] A. Traxler and E. Brewe. Equity investigation of attitudinal shifts in introductory physics. *Physical Review ST Physics Education Research*, 11:020132, Nov 2015.

[264] A. Traxler, R. Henderson, J. Stewart, G. Stewart, A. Papak, and R. Lindell. Gender fairness within the force concept inventory. *Physical Review Physics Education Research*, 14:010103, Jan 2018.

[265] A. L. Traxler, X. C. Cid, J. Blue, and R. Barthelemy. Enriching gender in physics education research: A binary past and a complex future. *Physical Review Physics Education Research*, 12:020114, Aug 2016.

[266] A. Uitto. Interest, attitudes and self-efficacy beliefs explaining upper-secondary school students' orientation towards biology-related careers. *International Journal of Science and Mathematics Education*, 12(6):1425–1444, 12 2014.

[267] B. Van Dusen and J. Nissen. Modernizing use of regression models in physics education research: A review of hierarchical linear modeling. *Physical Review Physics Education Research*, 15:020108, Jul 2019.

[268] P. Vincent-Ruz, K. Binning, C. D. Schunn, and J. Grabowski. The effect of math SAT on women's chemistry competency beliefs. *Chemistry Education Research and Practice*, 19(1):342–351, 2018.

[269] T. S. Volkwyn, B. Gregorcic, J. Airey, and C. Linder. Learning to use Cartesian coordinate systems to solve physics problems: The case of 'movability'. *European Journal of Physics*, 41(4):045701, Jun 2020.

[270] G. M. Walton and G. L. Cohen. A question of belonging: race, social fit, and achievement. *Journal of Personality and Social Psychology*, 92(1):82, 2007.

[271] G. M. Walton and T. D. Wilson. Wise interventions: Psychological remedies for social and personal problems. *Psychological Review*, 125(5):617, 2018.

[272] Z. H. Wan. Exploring the effects of intrinsic motive, utilitarian motive, and self-efficacy on students' science learning in the classroom using the expectancy-value theory. *Research in Science Education*, 2019.

[273] M.-T. Wang and J. Degol. Motivational pathways to STEM career choices: Using expectancy–value perspective to understand individual and gender differences in STEM fields. *Developmental Review*, 33(4):304–340, 2013.

[274] X. Wang. Why students choose STEM majors: Motivation, high school learning, and postsecondary context of support. *American Educational Research Journal*, 50(5):1081–1121, 2013.

[275] Y. Wang, L. Wang, H. Liang, D. Zollman, L. Zhao, and Y. Huang. Research on the small private online course (SPOC) teaching model incorporating the Just-in-Time Teaching (JiTT) method based on mobile internet for learning college physics. *European Journal of Physics*, 41(3):035701, Mar 2020.

[276] P. A. Westrick, H. Le, S. B. Robbins, J. M. Radunzel, and F. L. Schmidt. College performance and retention: A meta-analysis of the predictive validities of ACT scores, high school grades, and SES. *Educational Assessment*, 20(1):23–45, 2015.

[277] K. M. Whitcomb, Z. Y. Kalender, T. J. Nokes-Malach, C. Schunn, and C. Singh. Inconsistent gender differences in self-efficacy and performance for engineering majors in physics and other disciplines: A cause for alarm? In *Physics Education Research Conference 2019*, PER Conference, Provo, UT, July 24-25 2019.

[278] K. M. Whitcomb, Z. Y. Kalender, T. J. Nokes-Malach, C. Schunn, and C. Singh. Engineering students' performance in foundational courses as a predictor of future academic success. *International Journal of Engineering Education*, 36(4):1340–1355, 2020.

[279] H. Wickham. *tidyverse: Easily Install and Load the 'tidyverse'*, 2017. R package version 1.2.1.

[280] T. D. Wilson and P. W. Linville. Improving the performance of college freshmen with attributional techniques. *Journal of Personality and Social Psychology*, 49(1):287, 1985.

[281] P. Winkelman. Perceptions of mathematics in engineering. *European Journal of Engineering Education*, 34(4):305–316, 2009.

[282] P. H. Winne. Self-regulated learning. In J. D. Wright, editor, *International Encyclopedia of the Social & Behavioral Sciences*, volume 21, pages 535–540. Elsevier, Oxford, UK, 2 edition, 2015.

[283] E. B. Witherspoon and C. D. Schunn. Locating and understanding the largest gender differences in pathways to science degrees. *Science Education*, 104(2):144–163, 2019.

[284] B. Wong. Minority ethnic students and science participation: A qualitative mapping of achievement, aspiration, interest and capital. *Research in Science Education*, 46(1):113–127, 2016.

[285] D. S. Yeager and G. M. Walton. Social-psychological interventions in education: They're not magic. *Review of Educational Research*, 81(2):267–301, 2011.

[286] T. J. Yosso. Whose culture has capital? A critical race theory discussion of community cultural wealth. *Race Ethnicity and Education*, 8(1):69–91, 2005.

[287] N. T. Young and A. F. Heckler. Observed hierarchy of student proficiency with period, frequency, and angular frequency. *Physical Review Physics Education Research*, 14:010104, Jan 2018.

[288] C. Zabriskie, J. Yang, S. DeVore, and J. Stewart. Using machine learning to predict physics course outcomes. *Physical Review Physics Education Research*, 15:020120, Aug 2019.

[289] G. Zavala and A. Dominguez. Engineering students' perception of relevance of physics and mathematics. In *2016 ASEE Annual Conference and Exposition*. American Society for Engineering Education, Jun 2016.

[290] G. Zhu and C. Singh. Improving students' understanding of quantum measurement. II. Development of research-based learning tools. *Physical Review ST Physics Education Research*, 8:010118, Apr 2012.

[291] G. Zhu and C. Singh. Improving student understanding of addition of angular momentum in quantum mechanics. *Physical Review ST Physics Education Research*, 9:010101, Jan 2013.

[292] B. J. Zimmerman. Self-efficacy: An essential motive to learn. *Contemporary Educational Psychology*, 25(1):82–91, 2000.