THE UNIVERSITY OF CHICAGO

MATHEMATICS REQUIREMENTS, SELF-EFFICACY, AND THE GENDER GAP IN STEM

A BACHELOR THESIS SUBMITTED TO

THE FACULTY OF THE DEPARTMENT OF ECONOMICS

FOR HONORS WITH THE DEGREE OF

BACHELOR OF THE ARTS IN ECONOMICS

BY JESS XIONG

CHICAGO, ILLINOIS

MAY 2025

TABLE OF CONTENTS

ABSTRACT

This thesis investigates whether increasing high school math graduation requirements can reduce the gender gap in math self-efficacy among American students. Math self-efficacy, defined as students' belief in their ability to succeed in math, is a key predictor of advanced course-taking, STEM major selection, and long-term persistence. Using nationally representative data from the Educational Longitudinal Study of 2002, I apply a difference-in-differences design with propensity score matching to estimate the effects of moving from two to three and from three to four years of required math coursework. While the two-to-three-year change shows no effect, the shift to four years leads to statistically significant gains in self-efficacy, particularly after accounting for gender. Girls report consistently lower confidence than boys, even after controlling for academic ability, and further interaction analyses show that race and baseline math beliefs influence how students respond to increased requirements. These findings provide new causal evidence that structural policy can shape student confidence and suggest that graduation mandates may play a role in reducing gender disparities in math and STEM pathways.

*. I am deeply grateful to the people who made this thesis possible. First and foremost, I would like to thank my advisor, Professor Joensen, who also taught me Applied Microeconometrics. That class sparked my interest in economic research and laid the foundation for this thesis. Professor Joensen's guidance, from the early stages of the idea to helping me pivot after our data request fell through, was integral to every part of this project. I'd also like to thank Dylan Balla-Elliott, my TA for Applied Microeconometrics, who was not only an amazing TA but also continued to help me develop my thesis even after the course ended. Thank you for your support and for believing in my ideas. To my friends, thank you for supporting me throughout this process, for listening to my half-baked ideas, coming to my presentations, and encouraging me through the most stressful moments. I appreciate you more than you know. Lastly, I dedicate this thesis to every girl who was told she wasn't a "math person." I hope this work challenges that narrative and supports a broader rethinking of who belongs in math.

CHAPTER 1

INTRODUCTION

Women in the United States have made dramatic strides in educational attainment, surpassing men in college enrollment and degree completion. Yet despite these gains, they remain persistently underrepresented in science, technology, engineering, and mathematics (STEM). In 2020, women earned just 26 percent of bachelor's degrees in mathematics and computer science and held a similarly low share of jobs in those fields. These disparities reflect enduring gender inequalities in academic and professional spaces and contribute to wage gaps, occupational segregation, and the broader underutilization of talent in the STEM workforce.

What drives this underrepresentation? Traditional explanations point to performance differences or divergent preferences. But a growing body of research suggests another key driver: beliefs. High-achieving girls are often less confident in their math abilities than equally capable boys, and this confidence gap emerges early and deepens over time. As a result, many girls opt out of math-intensive coursework and career paths not because they lack ability, but because they underestimate it.

This thesis asks whether educational policy can intervene in this belief formation process. Specifically, can increasing high school math graduation requirements boost students' math self-efficacy, particularly for girls who may be less likely to enroll in advanced math voluntarily? Self-efficacy, or one's belief in their ability to succeed in a domain, is a powerful predictor of academic persistence, course selection, and entry into STEM. If policy can alter this internal perception, it may offer a lever for addressing disparities that have long resisted more conventional interventions.

To explore this question, I use data from the *Educational Longitudinal Study of 2002* (ELS:2002), a nationally representative panel that tracks students from 10th to 12th grade. I exploit variation in students' exposure to different state-level math graduation requirements by comparing cohorts subject to two, three, or four years of mandated math coursework. Because geographic identifiers are not available in the public-use dataset, I rely on students' self-reported requirements and implement a Difference-in-Differences design combined with Propensity Score Matching to improve covariate balance and identify causal effects.

The results reveal a striking asymmetry. Raising requirements from two to three years has no significant effect on self-efficacy, suggesting that modest curricular extensions may not meaningfully shift students' beliefs. But increasing the mandate from three to four years produces statistically significant gains in math self-efficacy, especially among students already on advanced tracks. These gains become more pronounced when gender is included in the model, indicating that average effects may obscure underlying gendered dynamics. Across all specifications, female students report significantly lower confidence than male peers, even after adjusting for academic performance. Interaction models reveal that these gender gaps are not static but evolve differently across race and baseline self-beliefs, highlighting the importance of an intersectional lens.

This thesis makes three core contributions. First, it offers causal evidence that high school graduation requirements can shape psychological outcomes, specifically students' confidence in their mathematical ability. Second, it demonstrates that curricular policy interacts with identity, producing heterogeneous effects by gender and race. Third, it advances methodological rigor in policy evaluation by pairing matching techniques with Difference-in-Differences in a setting where standard fixed effects cannot be applied.

More broadly, this study argues that educational policy does not merely allocate opportunities. It helps shape the internal narratives students hold about what they are capable of. For girls in particular, these narratives may be just as critical as skills in determining whether they see themselves in STEM. If we want more women to pursue technical fields, we must start not only by changing the pipeline but by reshaping the beliefs that guide who steps into it.

CHAPTER 2

LITERATURE REVIEW

In this section, I review three key areas of literature to support my hypothesis. First, I document descriptive evidence of the gender gap in STEM fields, outlining standard economic explanations, such as ability differences and financial incentives, and discussing their limitations. Second, I explore psychological explanations that highlight internal mechanisms like self-efficacy, identity, and mindset as critical determinants of educational and career decisions. Lastly, I examine empirical studies on policy interventions, particularly graduation requirements, and their impact on gendered academic trajectories and self-perceptions, arguing that structural mandates can serve as powerful levers for reducing gender disparities in STEM participation.

2.1 Economic Perspectives on STEM Gender Gaps

Numerous studies document women's underrepresentation in math-intensive majors and STEM occupations. The consensus is that the gender disparity is most pronounced in more mathematically intensive disciplines, such as engineering, computer science, physics, and economics, whereas women have reached parity or even majority in less math-intensive sciences such as biology or psychology [Ceci et al., 2014]. For example, women now earn about 57% of all bachelor's degrees in the U.S. but only 38–39% in STEM, with stark differences by field: women receive roughly 18–21% of computer science and engineering degrees but around 60% of degrees in biological sciences. These disparities also carry over into the workforce: In 2021, women made up only 18% of STEM workers [National Science Foundation, National Center for Science and Engineering Statistics, 2019].

Traditional economic research offers several potential explanations for women's underrepresentation in STEM fields, but these explanations appear to be only partially sufficient.

One common hypothesis attributes the disparity to ability-based differences, particularly at the upper end of the distribution. For example, Fryer and Levitt [2010] document that while there is no gender gap in math performance at school entry, a modest gap emerges in fifth grade and continues to widen into high school. The gap is most pronounced at the top of the distribution, where boys are more likely than girls to score in the highest percentiles. However, the authors find that this widening gap cannot be fully explained by observable factors such as family background, classroom environment, student behavior, etc. This suggests the presence of other mechanisms, potentially internal in nature, that cannot be measured through easily observable covariates.

One such mechanism may involve gender differences in how individuals psychologically respond to competitive or high-pressure contexts, which are often implicit features of STEM environments. Experimental work by Niederle and Vesterlund [2010] demonstrates that men are significantly more likely than women to opt into competitive compensation schemes, even when actual ability and performance are held constant. This suggests that gender differences in choice behavior may not stem from objective skill, but from subjective factors like confidence, anticipated success, and comfort with competition. The study thus highlights how internal psychological traits, such as mindset, risk tolerance, and self-assessed competence, can shape educational and career decisions in gendered ways.

In the context of STEM, which is often perceived as selective, evaluative, and high stakes, these preferences may lead women to opt out of certain courses or majors, not because they lack ability, but because they anticipate a lower fit or success in such settings. If advanced math and science classes are framed or experienced as competitive environments, women may be more likely to select out preemptively, even when they are equally capable. This behavioral pattern, grounded in psychological belief differences rather than cognitive gaps, offers a compelling explanation for why fewer women persist through the most math-intensive academic tracks and supports the broader view that the STEM gender gap reflects not just skill disparities, but identity, confidence, and mindset.

Labor market evidence adds another layer of complexity. Joensen and Nielsen [2015] show that for girls, completing more advanced and math-intensive coursework leads to significant earnings gains. These gains are often realized through entry into more competitive and higher paying careers. In contrast, there is no comparable marginal earnings effect for boys. This suggests that women who pursue advanced math reap greater economic returns, which makes their lower rates of selection into these fields puzzling from the perspective of rational choice theory. If women avoid high-return STEM pathways despite clear financial incentives, then standard economic models may be insufficient.

Together, these findings point to the importance of psychological factors in determining women's selection into STEM. Although some women may simply prefer other careers, others may be constrained by internalized doubts about their abilities or discomfort in high-stakes, male-dominated settings. Differentiating between these possibilities is important. Preferences should be respected, but when structural or psychological barriers discourage students from pursuing their potential, targeted interventions become necessary. Understanding how educational experiences shape these beliefs, particularly through math instruction and classroom environments, can help inform policies aimed at improving gender equity in STEM.

2.2 Psychological Mechanisms: Self-Efficacy and Academic Identity

A growing body of research suggests that psychological mechanisms, particularly self-efficacy and academic identity, are critical to understanding gender disparities in STEM. Self-efficacy, defined by Bandura [1977] as the belief in one's own ability to succeed in specific situations, strongly predicts academic persistence and achievement. In mathematics and science, girls consistently report lower self-efficacy than boys, even when achieving similar or higher grades. This "confidence gap", which often emerges in early childhood, significantly affects the selection of girls' courses and long-term participation in STEM pathways [Perez-Felkner et al., 2017].

According to Bandura [1977], four key sources shape self-efficacy:

1. **Mastery experiences**: Success builds confidence, yet girls are often steered away from early STEM challenges, limiting opportunities for positive reinforcement.

2. **Vicarious experiences**: Observing role models succeed can enhance belief in one's own abilities. The lack of visible women scientists and engineers weakens this channel.

3. **Social persuasion**: Encouragement strengthens confidence, but gender-based messages often signal that girls are less suitable for STEM, undermining self-belief [Eccles and Wang, 2016].

4. **Physiological states**: High anxiety, common among girls in math contexts, can be misinterpreted as inability, reducing self-perceived competence.

Since mathematics self-efficacy was shown by Perez-Felkner et al. [2017] to be a significant predictor of STEM career pursuit, policies which focus on the four sources of self-efficacy should be examined.

The expectancy-value theory developed by Eccles and Wang [2016] further contextualizes this dynamic. It posits that students' academic choices depend not only on their perceived competence in a subject but also on their comparative perceived abilities across different domains. Girls often exhibit higher verbal abilities compared to boys, and when faced with equivalent math skills but superior verbal skills, girls may disproportionately select out of math-intensive paths due to their comparative advantage in verbal areas. This reasoning underscores the importance of controlling for verbal self-efficacy in my analysis to accurately isolate the role of math self-efficacy and better understand the nuanced factors influencing girls' decisions to pursue or avoid STEM fields.

Interventions that target mindsets also play a significant role. A growth mindset, or the belief that ability can be developed with effort, has been shown to buffer against stereotype threat and increase resilience in challenging environments. [III and Banda, 2016] demonstrate that teaching girls to adopt growth mindsets improves math engagement in male-dominated settings. Similarly, grit, the tendency to maintain effort and interest over long periods, is positively associated with STEM persistence, but is most effective when aligned with strong self-efficacy and academic identity [III and Banda, 2016].

Because self-efficacy and academic identity so powerfully shape the educational trajectories of students, particularly for girls in math-intensive settings, this study focuses on key psychological constructs that emerge from the literature. These include math-specific self-efficacy, verbal self-efficacy, general self-concept under challenge, and beliefs about the growth mindset. By incorporating these variables into the analysis, the research captures a more nuanced understanding of how internal beliefs and orientations contribute to gendered patterns in STEM engagement.

2.3 Policy Levers and Interaction With Identity

My hypothesis is that graduation mandates can serve as institutional nudges that reshape students' academic trajectories by requiring rigorous coursework. I argue these structural interventions influence not only educational outcomes but also students' internal identity development, particularly by generating "forced" mastery experiences that enhance self-efficacy.

Joensen and Nielsen [2015] provide compelling evidence that lowering barriers to advanced coursework can significantly increase women's participation in technical and traditionally male-dominated programs. They document that women tend to opt out of math-intensive pathways early, often by the end of high school, highlighting that differential course taking behaviors begin well before college. Thus, policy interventions targeting high school coursework, such as math graduation mandates, can indirectly influence long-term outcomes by shaping students' perceptions of available academic options. If course selection is endogenous to students' confidence or perceived ability, compulsory advanced coursework could shift trajectories by preserving options for underconfident students who might otherwise prematurely self-select out. Hence, structural interventions that raise expectations could serve as critical levers for expanding participation in STEM fields.

Empirical studies reinforce this perspective. Goodman [2017] finds that increasing state math requirements significantly increased math course taking and subsequent earnings, notably among black students who previously enrolled less frequently in advanced math courses. These effects were concentrated among students with lower prior achievement, suggesting that the mandates effectively closed racial and achievement gaps by urging under-confident students to take on challenging coursework. Similarly, Cortes et al. [2015] showed that Chicago's 'double dose' algebra intervention, which doubled the instructional time in ninth grade algebra, substantially improved academic outcomes for students with initially weak math skills. The intervention provided structured mastery experiences through extended participation, bolstering the confidence and skills of the students.

Moreover, Jia [2021] shows that stricter high school math graduation requirements significantly increased STEM degree attainment, particularly for women. These mandates encouraged students to switch from non-STEM to STEM majors, underscoring that structural nudges can activate latent potential among traditionally underrepresented groups.

These findings collectively support the hypothesis that externally imposed academic rigor disproportionately benefits students with low academic confidence, particularly girls in math-intensive fields. By mandating challenging courses, schools facilitate mastery experiences that build self-efficacy and reshape students' academic identities. My hypothesis is that these effects operate through Bandura's four foundational mechanisms of self-efficacy. First, mandates create mastery experiences by steering students toward early STEM challenges

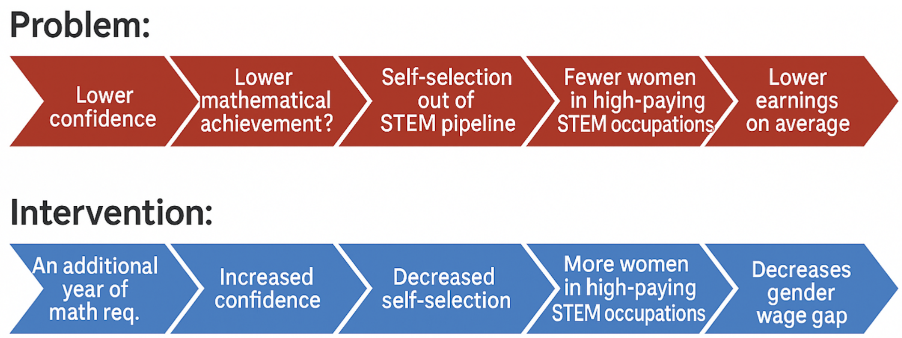


Figure 2.1: Hypothesized pathway linking math graduation requirements to gender equity in STEM

they might otherwise avoid, allowing confidence to develop through successful completion. Second, increased female enrollment in advanced math promotes vicarious experiences, as girls observe peers who resemble them succeeding, thereby normalizing participation and expanding the pool of role models. Third, compulsory coursework reduces vulnerability to negative social persuasion: Students cannot be discouraged or dissuaded from engaging in challenging topics. Finally, although more speculative, repeated exposure to math-intensive environments can alleviate physiological anxiety, helping students become more comfortable with quantitative subjects. Collectively, these mechanisms suggest that well-designed graduation mandates can promote equity by nurturing internal confidence crucial for sustained STEM engagement.

CHAPTER 3

DATA

3.1 Data Source

This study uses data from the *Educational Longitudinal Study of 2002* (ELS:2002), a nationally representative panel survey conducted by the National Center for Education Statistics (NCES). The study follows a cohort of approximately 14,200 students who were in 10th grade in 2002, with follow-up waves in 2004 and 2006. In addition to student responses, ELS:2002 includes information collected from parents, teachers, and school administrators, as well as high school transcript records, providing a rich view of students' educational experiences and postsecondary transitions.

To prepare the data for analysis, I selected variables capturing students' math self-efficacy, academic performance, and demographic background. Observations with missing or suppressed responses on key covariates or outcomes were excluded. I also removed variables that were fully suppressed by the data provider. Categorical variables such as gender, race, parental education, and income were recoded and relabeled for clarity in summary and regression tables.

Two analytic samples were constructed based on students' exposure to different state-level math graduation requirements. The first compares students in states requiring three versus four years of high school math (*3v4 cohort*), while the second compares those in states requiring two versus three years (*2v3 cohort*). Each sample includes a binary treatment indicator to capture assignment to the higher-requirement condition. For each cohort, the data were reshaped to long format to estimate models using a Difference-in-Differences (DiD) framework, and a binary indicator was created to distinguish between baseline (2002) and follow-up (2006) observations.

3.2 Variable Definition and Summary Statistics

3.2.1 Self Perception and Other Psychological Variables

In my analysis, I construct control variables for general perseverance, verbal self-efficacy, and math growth mindset alongside the outcomes of 10th- and 12th-grade math self-efficacy. These constructs capture different dimensions of motivation in line with psychological theory that could confound observed changes in math-specific self-efficacy.

Math self-efficacy refers to a student's belief in their ability to perform specific math-related tasks, such as succeeding on tests or mastering material. As shown in previous literature, it directly influences effort, persistence, and engagement in math learning. In ELS:2002, math self-efficacy is a variable constructed by the survey administrators that was measured through an index based on students' ratings of how often the following statements apply to them: (a) I'm confident that I can do an excellent job on my math tests, (b) I'm certain I can understand the most difficult material presented in math texts, (l) I'm confident I can understand the most complex material presented by my math teacher, (r) I'm confident I can do an excellent job on my math assignments, and (u) I'm certain I can master the skills being taught in my math class.

To isolate math-specific confidence from broader perseverance tendencies, I follow Perez-Felkner et al. [2017]'s methodology and construct a general perseverance index based on students' responses to the following items: (e) When I sit myself down to learn something really hard, I can learn it, (j) When studying, I try to work as hard as possible, (o) When studying, I keep working even if the material is difficult, (s) When studying, I try to do my best to acquire the knowledge and skills taught, and (v) When studying, I put forth my best effort.

It is necessary to control for Verbal Self-Efficacy because Perez-Felkner et al. [2017] showed that girls with both high math and high verbal self-efficacy are more likely to select out of STEM fields compared to those with high math but moderate verbal ability. Thus, it is necessary to control for verbal ability to account for comparative academic self-concept effects. I construct a Verbal Self-Efficacy index based on students' responses to: (c) I'm certain I can understand the most difficult material presented in English texts, (f) I'm confident I can understand the most complex material presented by my English teacher, and (m) I'm certain I can master the skills being taught in my English class.

Finally, in line with mindset theory, I control for math growth mindset, measured by student agreement with (a) Most people can learn to be good at math. Students with a growth mindset are more likely to persist through challenges, while those who believe intelligence is innate (fixed mindset) are less likely to attempt or succeed in difficult tasks. Girls are more likely to have a fixed mindset, making it especially important to adjust for this factor when examining gendered patterns in math self-efficacy.

Table 3.1: Ability Perception Variables

c

Type	**Variable**	
Control	General Perseverance Index	Constructed from five survey items (BYS89E, BYS89J,
Control	Verbal Self-Efficacy Index	Constructed
Outcome	Math Self-Efficacy Index (10th Grade)	The Base Year Math Self-Efficacy variable (BYMATHSE
Outcome	Math Self-Efficacy Index (12th Grade)	The First Follow-up Math Self-Ef
Outcome	Growth Mindset Measure	Based on a single surv

Notes: Variables adapted following the constructions described by Perez-Felkner, Nix, and Thomas (2017).

3.2.2 Demographic Control Variables

Given the constraints of the public-use ELS:2002 dataset, I rely on a reduced set of theoretical control variables to proxy for students' academic preparedness and demographic background. Many ideal controls, such as 10^{th} grade GPA, are not available in the public data set. Grade-specific GPA is restricted, and the only accessible alternative is cumulative GPA, which is both categorical (e.g., 1.0–1.9, 2.0–2.9, etc.) and not pre-treatment, as it incorporates post-treatment information from grades beyond 10^{th} grade. Because cumulative GPA could bias causal inference by capturing treatment effects, I exclude it from analysis.

Instead, I use base-year math and verbal test scores as pre-treatment academic controls. These are Item Response Theory (IRT) estimated number-right scores, designed to reflect a student's underlying ability rather than raw performance alone. Specifically, the BYTXMIRR variable represents the estimated number of math items a student would have answered correctly out of 72 questions in the ELS:2002 math item pool had they attempted all items. Likewise, BYTXRIRR represents the analogous reading score based on 51 reading items. These scores are also computed using IRT. The probability of a correct response is estimated using logistic functions calibrated to item-level parameters, and these probabilities are then summed to produce a continuous expected score.

IRT has several advantages over raw test scores. It adjusts for the difficulty of each item, allows for comparisons across students even if they received different item sets, and provides more accurate ability estimates when test lengths vary. This makes the IRT-estimated number-right scores a strong proxy for cognitive ability in the absence of detailed GPA data.

I also intended to control for regional and school-level fixed effects, but all geographic identifiers in ELS:2002, including state, county, census region, and school codes, are restricted-use and not accessible in the public data. As such, I cannot account for variation due to local policy environments, quality of instruction, or school-level peer effects. This is a limitation of the current analysis, although I attempt to mitigate it by incorporating survey weights and controlling for observable demographic characteristics, such as race/ethnicity, parental education, household income, and urbanicity.

Table 3.2: Summary of Demographic Control Variables

c

Variable	
Gender	
Race/Ethnicity	C
Parent Education	
Household Income	
Urbanicity	
Base Year Math Score (10th Grade)	Math IRT-estimated number right score, representing the expected r
Base Year Verbal Score (10th Grade)	ELS:2002 Reading IRT-estimated number right score, representing the ex

Notes: Race/ethnicity, parent education, household income, and urbanicity were categorical variables and were either dummy-coded or treated as ordinal depending on the analysis.

3.2.3 Summary Statistics

After applying all cleaning and filtering procedures, the final analytic samples include 4,690 students in the 3v4 cohort and 5,230 students in the 2v3 cohort. All regression models apply the longitudinal panel weights provided by NCES to ensure nationally representative estimates. Descriptive statistics showing the difference of each analytic sample (2, 3, and 4 years of math required) from the mean are shown below.

Table 3.3: Demographic Summary by Treatment Group

Category	Variable	Full Sample (%)	2-Year−Full	3-Year−Full	4-Year−Full
Gender	Male	50.2600	-0.6674	-0.9657	-1.8983
	Female	49.7400	0.6674	0.9657	1.8983
Race/Ethnicity	White	62.1929	11.9316***	0.5947	-0.2814***
	Black	13.7378	-9.5726***	-0.7028	6.9153***
	Hispanic	15.1018	-2.4924***	0.1341	-5.1280***
	Asian/PI	3.8693	0.6449***	0.1781	-0.3830***
	AIAN	0.9610	0.0029***	0.0886	-0.5421***
	Multi-race	4.1373	-0.5145***	-0.2927	-0.5807***
Parent Educ.	HS or less	30.2911	-3.9096	-0.6371	-7.3821**
	$<$ 4-yr degree	26.3794	1.4072	-1.2896	-1.0246**
	Bachelor	25.0033	1.3736	0.9272	4.3473**
	$>$ Bachelor	18.3263	1.1288	0.9995	4.0595**
HH Income	$\leq$ \$35K	33.1989	-4.0522*	-2.4322	-4.7937***
	\$35–50K	19.8209	-0.0657*	-0.0690	-2.5844***
	\$50–100K	34.1413	3.7180*	1.7117	1.9879***
	$\geq$ \$100K	12.8388	0.3999*	0.7896	5.3902***
Urbanicity	Urban	30.0100	-8.4695***	-1.3563	7.2116***
	Suburban	50.2434	8.1833***	1.0414	-9.5327***
	Rural	19.7466	0.2862***	0.3149	2.3211***
Student Ability	Math Score	37.5253009	2.0848861	1.0338096	2.0606275
	Verbal Score	29.6536123	1.2385990	0.9009506	1.5083879

Notes: *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$ — *stars mark the difference between the treatment group and the full-sample mean.*

Table 3.4: Outcome Summary by Treatment Group

cc

Category	Outcome	Full (mean)	2-Year-Full	3-Year-Full	4-Year-Full
Non-Math Measures	General Perseverance	2.74	0.01	0.04*	0.14***
	Verbal Self-Efficacy	2.66	0.02	0.03	0.05
Math Measures	Math Self-Efficacy (10th Grade)	2.47	0.02	0.02	0.16***
	Math Self-Efficacy (12th Grade)	2.52	-0.02	-0.02	0.13***
	Growth Mindset (10th Grade)	2.05	0.00	-0.01	0.00

Notes: $^{***}\, p < 0.001$; $^{**}\, p < 0.01$; $^{*}\, p < 0.05$ *— stars mark the difference between the treatment group and the full-sample mean.*

Table 3.4 shows that higher math requirements are associated with improvements in non-cognitive outcomes, particularly math-specific self-efficacy. Students in the 2-year requirement group show no meaningful differences from the full-sample mean. In the 3-year group, general perseverance is slightly higher, suggesting a small gain in overall academic persistence. The most notable effects appear in the 4-year requirement group, where general perseverance, 10th Grade math self-efficacy, and 12th Grade math self-efficacy are all significantly higher. These results support the hypothesis that requiring more years of math may improve students' confidence in their mathematical abilities and their willingness to persist through academic challenges, especially since unrelated domains such as verbal self-efficacy or growth mindset are unchanged.

3.3 Preliminary Patterns

3.3.1 Math Course-Taking Behavior Responds to Policy Incentives

Before turning to causal analysis, it is important to understand how student math course-taking behavior varies in different policy environments. This analysis operates under the assumption that students generally progress through math courses in a sequential order—moving from Algebra I to Geometry, then Algebra II, and finally to Trigonometry, Pre-Calculus, or Calculus—consistent with the typical structure of the American high school curriculum. Students typically begin high school in Algebra I, meaning that a 2-year requirement usually covers Algebra I and Geometry, a 3-year requirement adds Algebra II, and a 4-year requirement pushes students into more advanced courses such as Trigonometry, Pre-Calculus, or Calculus.

Table 3.5 looks at the distribution of students' highest math course taken across states that require 2, 3, or 4 years of high school mathematics. This is visualized in Figure 3.1. The results suggest that students in states with 4-year math requirements are significantly more likely to reach advanced courses, with 64.3 percent listing Trigonometry, Pre-Calculus, or Calculus (henceforth referred to as Trig+) as their highest course taken. In contrast, only 43.1 percent of students in 2-year states report reaching this level, with higher proportions remaining in lower-level courses such as Pre-Algebra, Algebra I, and Geometry, aligning with the hypothesis that stricter requirements may shift students into higher levels of math.

Table 3.5: Highest Math Course Taken By Years of Math Required

Highest Math Course	**2 Years**	**3 Years**	**4 Years**
No math course	1.1	0.8	0.8
Pre-algebra/general math	5.7**	3.1*	1.5**
Algebra I	7.6**	4.3*	2.3**
Geometry	13**	10.6	8.2**
Algebra II	28.4	30**	21.8**
Trig, Pre-calc, or Calculus	43.1**	50.3	64.3**
Other/Unknown	1.2	0.9	1.1

Note: Asterisks indicate statistically significant deviations based on standardized residuals from a chi-squared test of independence. * indicates residual > 2 ($p<0.05$); ** indicates residual > 3 ($p<0.01$).

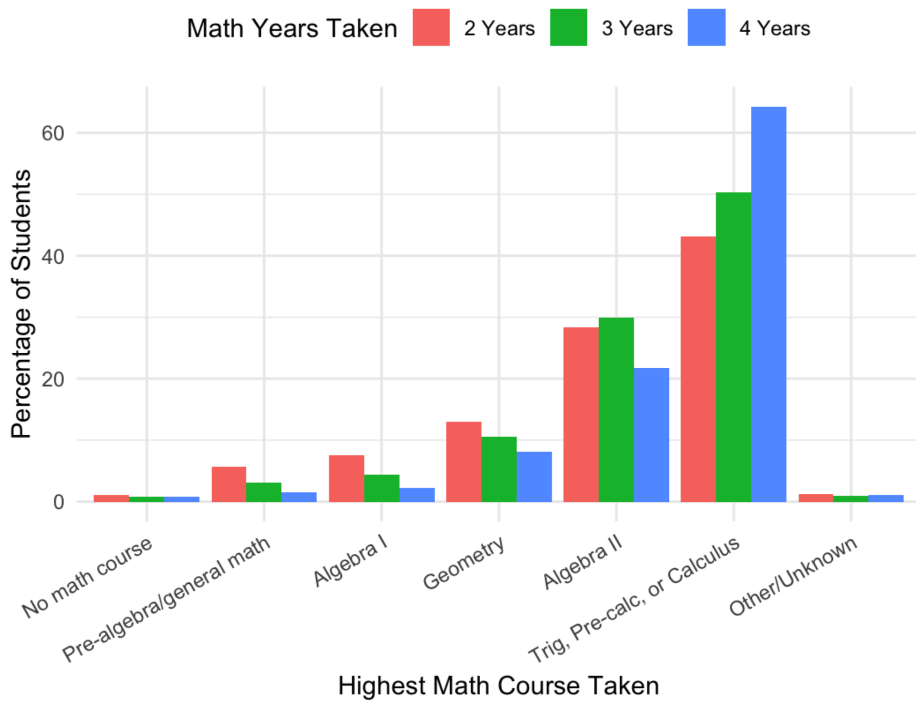


Figure 3.1: Highest Math Course Taken By Years of Math Required

In order to verify this more formally, I run a first-stage regression to estimate how being in each treated state changes the probability that the student enrolls in more advanced math courses. I modeled this using an ordered logistic regression where the dependent variable was the highest level of math a student took, and the main predictor was the number of math years required, controlling for all demographic covariates. The results in Table 3.6 show that in the 2-year to 3-year requirement comparison, there is a decline in predicted enrollment in lower-level courses below Trig+, accompanied by a notable increase in the probability of taking Trigonometry or higher, from 35.7% to 48.1%. This suggests that the additional required year of math is not simply being spent repeating or lingering in existing coursework. Instead, many students advance further, particularly those who reached Algebra II by Grade 11. This finding is especially striking because a 3-year requirement should, in principle, be pushing more students to reach Algebra II, the natural third-year course in the sequence. While this is indeed occurring, the data suggest that an even larger share of students are progressing beyond Algebra II into more advanced coursework. One possible explanation is that many high-achieving students had already reached Algebra II by Grade 11 and would have otherwise stopped there. By mandating a third year, the policy effectively requires these students to enroll in Trigonometry or Pre-Calculus in Grade 12, thus accelerating their math trajectory. In this way, the reform may disproportionately affect students at the upper end of the ability distribution, nudging them into college-preparatory tracks they might not have chosen voluntarily.

The effect becomes even more prominent in the 3-year to 4-year requirement comparison. The predicted probability in Trig+ enrollment increases from 49.2% to 60.9%, while those for all lower courses decline. This pattern suggests that the fourth required year is not displacing course-taking at lower levels but instead is likely facilitating continued progression for students already on a college-preparatory path. These students, having completed Algebra II by Grade 11, now appear more likely to pursue advanced math as seniors.

Overall, the policy not only increases minimum attainment but also accelerates the trajectory of students who were already ahead, generating a cascading effect up the math course ladder. This upward shift highlights the heterogeneity in how students respond to increased requirements: while the policy's minimum target may be Algebra II, its largest effects may be concentrated among students at the upper margin who are now nudged into more advanced coursework.

Table 3.6: Predicted Probabilities of Taking Each Math Course Level

Group	No Math	Pre-Algebra	Algebra I	Geometry	Algebra II	Trig+
Panel A: 2-Year vs. 3-Year Requirement						
Control	0.004	0.021	0.040	0.125	0.453	0.357
Treated	0.003	0.013	0.025	0.083	0.396	0.481
Panel B: 3-Year vs. 4-Year Requirement						
Control	0.003	0.012	0.021	0.082	0.391	0.492
Treated	0.002	0.008	0.013	0.054	0.314	0.609

3.3.2 Higher-Level Math is Skewed by Prior Ability and Advantage

A student's prior math ability is a strong predictor of the level of math they ultimately complete. As shown in Figure 3.2 and Table 3.7, there is a clear upward trend in base-year math scores by highest math course taken. Students who completed more advanced math courses had higher average base-year scores, with mean scores increasing from 26.5 among students whose highest course was Pre-algebra to 44.8 among those who reached Trigonometry, Pre-Calculus, or Calculus. These differences are statistically significant beginning with Geometry ($p < 0.01$), and the standard deviations indicate meaningful variation within each group. The strong relationship between prior ability and math course-taking justifies the inclusion of base-year math score as a control variable in the analysis.

Table 3.7: Base Year Math Score by Highest Math Course Taken

Highest Math Course	**Mean Math Score (SD)**	**n**
No math course	28.0 (11.0)	133
Pre-algebra/general math	26.5 (8.6)	597
Algebra I	28.3 (8.4)	900
Geometry	31.7 (9.3)**	1861
Algebra II	35.7 (9.7)**	4223
Trig, Pre-calc, or Calculus	44.8 (10.5)**	6821
Other/Unknown	32.9 (11.6)**	1357

Note: Asterisks indicate statistically significant differences compared to the "No math course" group based on Tukey's Honest Significant Difference (HSD) test. * indicates $p < 0.05$; ** indicates $p < 0.01$.

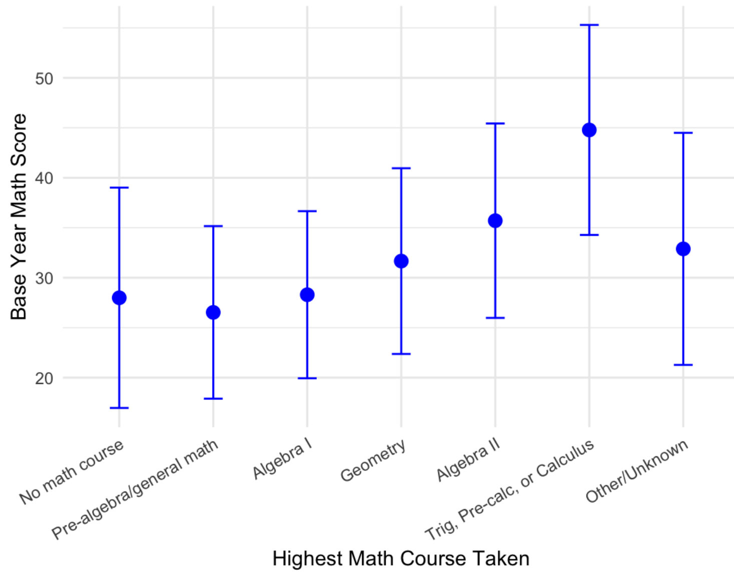


Figure 3.2: Base Year Math Score by Highest Math Course Taken

As expected, academic preparation is not evenly distributed acoss the student population. Figures 3.3 and 3.4 reveal strong socioeconomic gradients in math course-taking. Students from higher-income households and those with more highly educated parents are disproportionately represented in advanced math courses like Trig+. In contrast, students who stop at lower levels, such as Pre-algebra or Algebra I, are more likely to come from low-income families and households where parents did not complete college. These structural disparities suggest that course-taking reflects not only prior academic skill, but also access to resources, support, and expectations that shape long-term educational trajectories.

Taken together, these patterns highlight the importance of accounting for both individual academic readiness and socioeconomic context in the empirical analysis. Controlling for base-year math scores, family income, and parental education allows the model to isolate the effect of math requirements from these preexisting sources of variation.

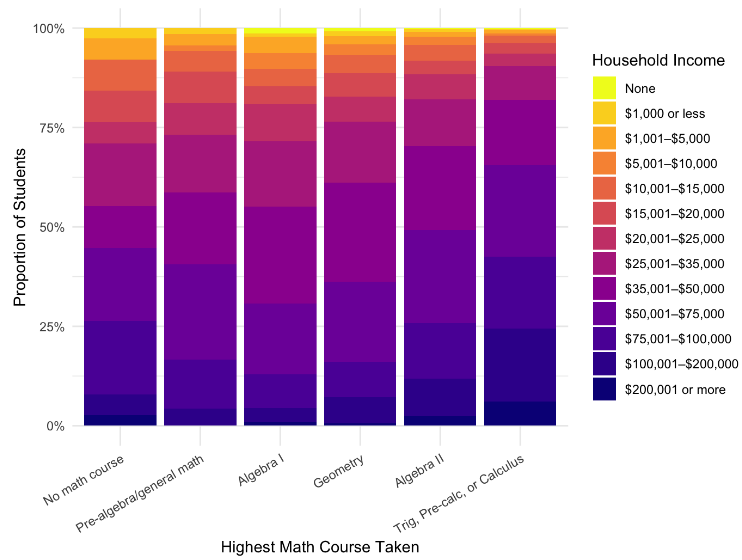


Figure 3.3: Household Income by Highest Math Course Taken

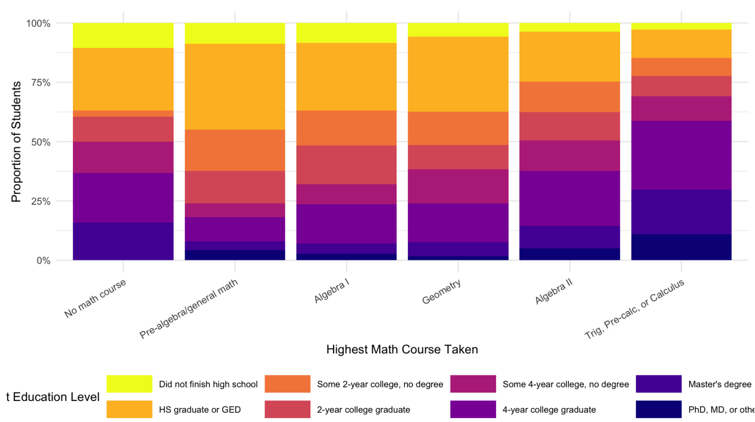


Figure 3.4: Parent Education by Highest Math Course Taken

3.3.3 Gendered Sorting Patterns Emerge

Table 3.8 reveals a striking reversal in gender representation as students progress through the high school math curriculum. At first glance, it appears that women outnumber men in most math courses. However, this pattern must be interpreted in the context of the overall composition of the sample: 54.1% female and 45.9% male. To assess whether each course exhibits a disproportionate representation of either gender, I conducted a two-sided binomial test comparing observed gender shares to these baseline proportions. The results in Table 3.8 reveal that male students are significantly overrepresented in Geometry (50.6% vs. expected 45.9%), while female students are significantly overrepresented in Algebra II (56.5% vs. expected 54.1%). Although women make up a slight majority in trigonometry, pre-calculus, and calculus (53.9%), this share does not differ significantly from expectations. These findings suggest that while female students generally persist at higher rates through the math pipeline, the most pronounced gender sorting occurs in the transition from mid-level to upper-level coursework, particularly at the Algebra II stage, which is often a critical juncture for STEM preparation.

Table 3.8: Gender Representation by Highest Math Course

Math Course	**Male (%)**	**Female (%)**
No math course	40.5 (−5.4)	59.5 (+5.4)
Pre-algebra/general math	48.5 (+2.6)	51.5 (−2.6)
Algebra I	49.6 (+3.6)	50.4 (−3.6)
Geometry	50.6 (+4.6**)	49.4 (−4.6**)
Algebra II	43.5 (−2.4**)	56.5 (+2.4**)
Trig, Pre-calc, or Calculus	46.1 (+0.2)	53.9 (−0.2)
Other/Unknown	51.3 (+5.3)	48.7 (−5.3)

*A two-sided binomial test was used to assess whether the observed gender proportion in each course significantly differed from the overall sample average (Female = 54.1%, Male = 45.9%). Asterisks denote significance: ** indicates* $p < 0.05$.

3.3.4 Psychological Patterns: Math Self-Efficacy Trajectories Diverge

Figures 3.5 through 3.7 illustrate patterns in math self-efficacy over time by gender, math ability decile, and highest math course taken. Figure 3.5 shows that across all ability deciles, male students report higher math self-efficacy than female students in both 10th and 12th grades. Figure 3.6 further illustrates that while both genders show variability in changes to self-efficacy from 10th to 12th grade, females tend to experience either greater improvements or smaller declines in the middle deciles, whereas high-ability female students (Q10) show the largest decrease in self-efficacy.

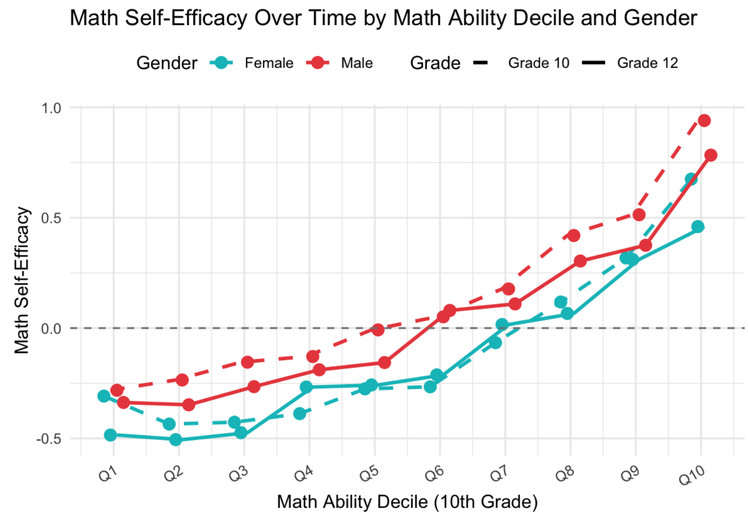


Figure 3.5: Math Self-Efficacy Over Time by Math Ability Decile and Gender

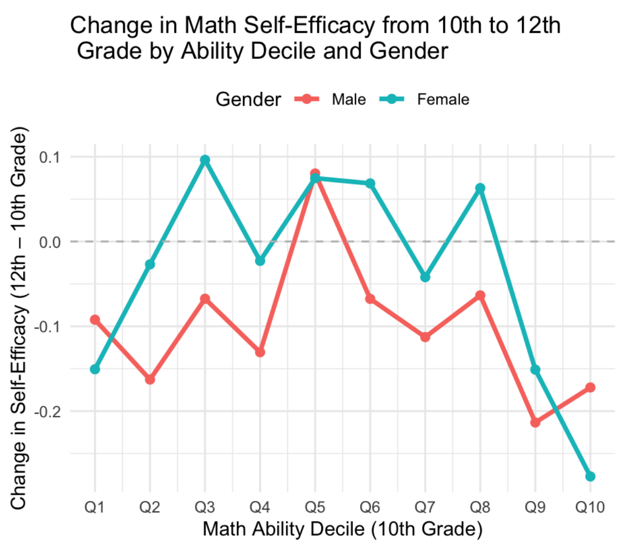


Figure 3.6: Change in Math Self-Efficacy from 10th to 12th Grade by Ability Decile and Gender

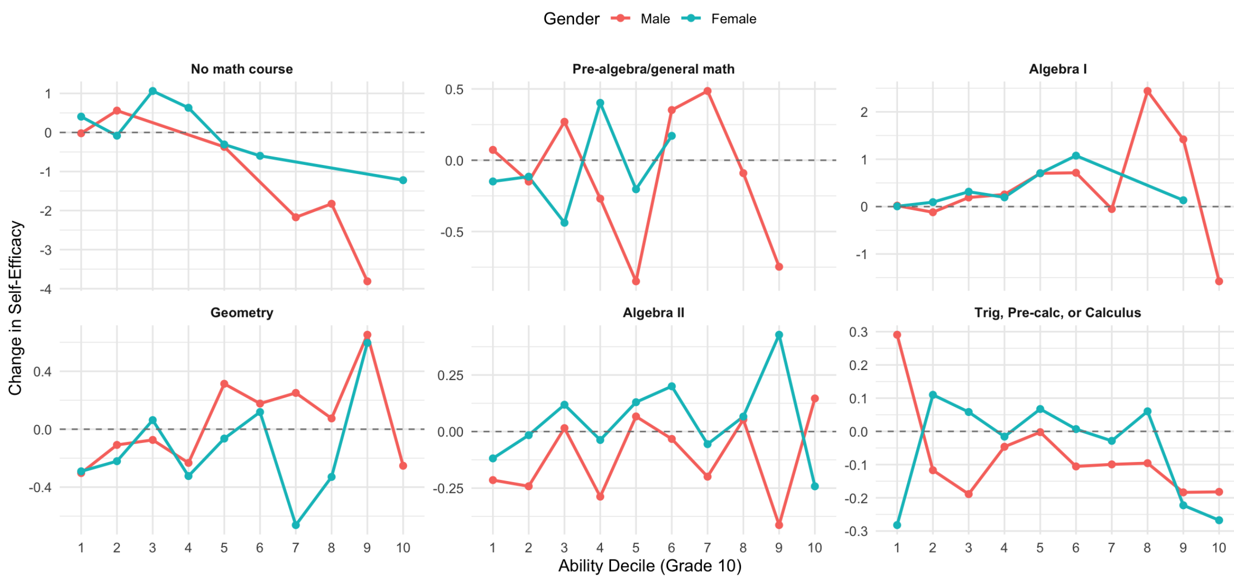
Figure 3.7 examines changes in self-efficacy disaggregated by students' highest math course taken, gender, and ability decile in 10th Grade. Among students who reach Algebra II or Trigonometry/Pre-Calculus/Calculus, female students generally experience positive changes in self-efficacy, suggesting that exposure to higher-level coursework may reinforce confidence, especially for those in mid-range ability deciles. Notably, however, this trend reverses at the very top of the ability distribution: in the highest deciles (Q9–Q10) of the Trig+ group, women's self-efficacy drops below that of men. This could reflect increased pressure or threat to stereotypes in advanced classrooms, where expectations and comparison stakes are highest.

In contrast, changes in self-efficacy are far more erratic in Pre-Algebra/General Math and Algebra I. These courses show high variability across deciles, with some large swings and missing data points, particularly for female students. This volatility limits interpretability, but may point to inconsistent instructional environments or unstable student confidence in lower-track settings.

In Geometry, girls show a notable dip in self-efficacy in the middle deciles, especially relative to boys, before catching up again by the ninth decile. However, missing data in some high deciles again makes it difficult to draw firm conclusions at the upper end.

These patterns suggest that course level moderates the trajectory of self-efficacy differently by gender. Upper-level math can stabilize or even improve girls' confidence, especially in the middle of the ability distribution, but this benefit is not uniform. For the highest-achieving girls, self-efficacy appears more fragile, even in the most advanced classrooms.

Figure 3.7: Change in Math Self-Efficacy from 10th to 12th Grade by Course, Gender, and Ability Decile



Note: Missing points indicate no students of that gender and ability decile took the given course.

Lastly, I verify the idea that higher mathematical self-efficacy increases the likelihood of choosing a STEM career. To analyze how math self-efficacy influences the intended field of study of the students, I grouped the declared majors into six broad categories. STEM includes math, engineering, physical sciences, and computer science. Applied professions capture practice-oriented, but often STEM-focused fields such as healthcare, nursing, and premed tracks. Humanities/Social Science, Vocational/Other, and Undecided are retained as is. I added Environmental Studies as a separate category, as it may refer to a variety of studies and it is unclear if it is in STEM.

Using a multinomial logistic regression, I estimated how math self-efficacy in 12th grade predicts the probability of selecting each major category, controlling for academic and demographic characteristics. The results, visualized in Figure 3.8. In the multinomial logit model, math self-efficacy significantly increases the likelihood of selecting a STEM or Applied Professions major, with no significant association for Environmental Studies, Undecided, or

Vocational/Other. A table of coefficients can be found in the Appendix.

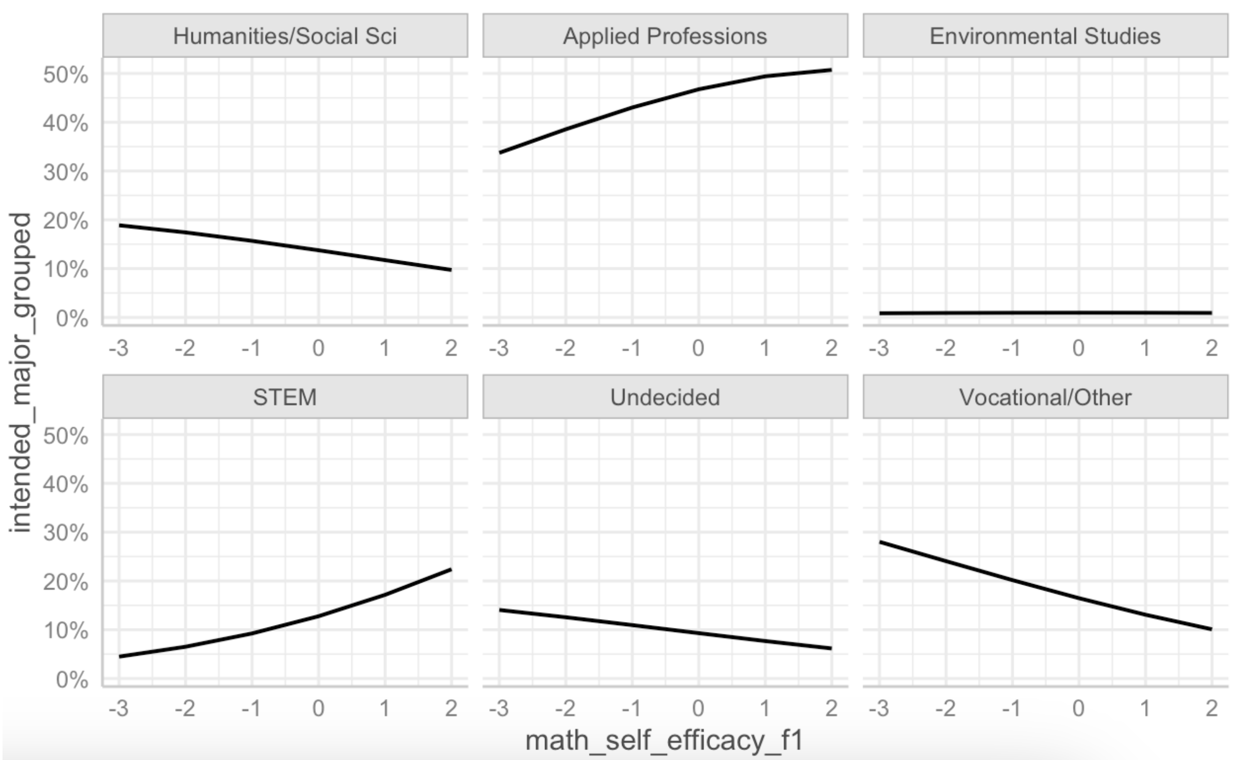


Figure 3.8: Predicted Probability of Intended Major By 12th Grade Math Self-Efficacy

CHAPTER 4

METHODOLOGICAL FRAMEWORK

4.1 Research Motivation

In this paper, I intend to test whether state-level increases in high school math graduation requirements have a causal effect on students' math self-efficacy, with a particular focus on whether these effects are stronger for girls. The intuition is that girls, who tend to perform well in math but often underestimate their abilities, may benefit from increased exposure to math coursework. If higher requirements reduce the option to self-select out of math, they may ultimately bolster confidence and persistence.

Initially, this study intended to employ a staggered difference-in-differences (DiD) approach, leveraging state-level variation in high school math requirements. This design would have provided a robust causal framework to measure the impact of mandatory math coursework on students' mathematical self-efficacy and subsequent STEM retention. However, due to unforeseen data access limitations—specifically being unable to obtain geocodes identifying students' states—this original methodological approach could not be implemented. Consequently, the empirical strategy was adjusted accordingly, as detailed below. A detailed description of the original staggered DiD approach is provided in the conclusion.

4.2 Empirical Strategy

To estimate the causal effect of more rigorous high school math graduation requirements on students' self-perceived math ability, I implement a Difference-in-Differences (DiD) design. This approach compares changes in math self-efficacy over time between students exposed to different levels of state-mandated math coursework. Although geographic identifiers are not available in the restricted-use ELS:2002 dataset, the survey includes a variable that records the number of math years required for graduation in each student's state of residence. I use this variable to define treatment status: students in states with more stringent math requirements constitute the treatment group, while those in states with less demanding requirements serve as the control group.

Because state-level fixed effects cannot be directly controlled for in the absence of geographic identifiers, I employ Propensity Score Matching (PSM) prior to estimation to improve balance between treatment and control units. As Lee et al. [2025] notes, combining PSM with a Difference-in-Differences framework can strengthen causal inference by reducing bias from observable confounders, particularly when randomized designs or geographic identifiers are unavailable. This helps ensure that comparisons are made between students who are similar in observed characteristics and likely to have been exposed to higher requirements.

My analysis focuses on two distinct policy contrasts: first, comparing students in states requiring two versus three years of math coursework (the 2v3 cohort); and second, comparing those in states requiring three versus four years (the 3v4 cohort). In each comparison, the treatment group includes students from states with the higher requirement, while the control group includes those from states with the lower requirement. The outcome variable is math self-efficacy, measured at two points in time: the base year (2002) when students were in 10th grade, and the first follow-up (2006) when students were in 12th grade.

To estimate both the overall effect of more stringent math graduation requirements and any differential effects by gender, I estimate two DiD models using separate matching specifications. The first model is designed to capture the average treatment effect across all students. In this specification, I include gender as a covariate in the propensity score matching (PSM) procedure to ensure treated and control students are balanced on baseline gender composition. This helps isolate the overall policy impact without confounding from gender differences in group composition.

To investigate whether the policy had heterogeneous effects on male and female students, I estimate a second model that includes a triple interaction term between treatment status, time, and gender. Importantly, for this model, I exclude gender from the set of covariates used in the matching procedure. This is a necessary adjustment because including gender in the matching process forces the treated and control groups to be balanced on gender by design, effectively removing any variation in gender composition between groups and preventing the estimation of gender-specific treatment effects. By excluding gender from the matching step, I preserve the natural gender distribution within treatment and control groups, allowing for valid identification of gender heterogeneity through interaction terms.

These two specifications together enable me to answer complementary questions: (1) What is the average effect of more rigorous math graduation requirements on students' math self-efficacy? and (2) Does this effect differ by gender? The first model provides a population-level estimate of the policy impact, while the second uncovers whether the policy amplified or narrowed gender disparities in students' math-related self-beliefs.

The DiD framework estimates the following linear model to identify the average treatment effect:

$$\text{SelfEfficacy}_{it} = \alpha + \beta_1 \text{Treated}_i + \beta_2 \text{Post}_t + \beta_3 (\text{Treated}_i \times \text{Post}_t) + \mathbf{X}_i' \boldsymbol{\gamma} + \epsilon_{it} \tag{4.1}$$

where SelfEfficacy_{it} is the math self-efficacy of student i at time t, Treated_i is a binary indicator equal to 1 if student i was subject to the more stringent math graduation requirement, and Post_t is a time indicator equal to 1 for the 2006 follow-up and 0 for the 2002 base year. The interaction term $\text{Treated}_i \times \text{Post}_t$ captures the DiD estimate of interest, denoted by β_3. The vector $\mathbf{X}_i$ includes baseline covariates such as parental education and school urbanicity. The error term ϵ_{it} captures unobserved factors affecting self-efficacy.

To investigate gender-specific treatment effects, I estimate a second model that includes a triple interaction between treatment status, time, and gender:

$$
\begin{aligned}
\text{SelfEfficacy}_{it} = \alpha &+ \beta_1 \text{Treated}_i + \beta_2 \text{Post}_t + \beta_3 \text{Female}_i \\
&+ \beta_4 (\text{Treated}_i \times \text{Post}_t) + \beta_5 (\text{Treated}_i \times \text{Female}_i) \\
&+ \beta_6 (\text{Post}_t \times \text{Female}_i) + \beta_7 (\text{Treated}_i \times \text{Post}_t \times \text{Female}_i) \\
&+ \ \mathbf{X}_i' \boldsymbol{\gamma} + \epsilon_{it}
\end{aligned}
\tag{4.2}
$$

In this specification, Female_i is a binary indicator for whether student i is female. The coefficient β_7 on the triple interaction term identifies the differential effect of the policy on math self-efficacy for female students, relative to their male counterparts. The other interaction terms serve to flexibly control for any gender-specific time trends or baseline differences in treatment effects.

4.3 Propensity Score Matching

A key identification assumption in DiD models is that in the absence of treatment, the treatment and control groups would have followed parallel trends, so self-efficacy would have changed in a similar manner. However, due to data limitations, only a single pre-treatment observation (10th Grade Math Self-Efficacy) is available, precluding direct tests of this assumption.

To mitigate concerns about non-parallel trends, Propensity Score Matching (PSM) is used to pre-process the data. Matching improves covariate balance and reduces reliance on model-based extrapolation. As argued in Stuart [2010], combining matching with DiD helps make the parallel trends assumption more plausible when pre-treatment outcomes are unavailable. With a non-random assignment to treatment, there is always the concern that the treatment states would have followed a different trend than the control states. However, if one can control for the factors that differ between the groups and that would lead to differences in time trends (and if these factors are exogenous), then the true effect from the treatment can be estimated. In this context, matching based on pre-treatment covariates serves to balance observed characteristics that may be correlated with both the likelihood of treatment and the potential outcomes. This helps simulate the counterfactual scenario in which treated units are statistically similar to untreated units in the absence of treatment. In its simplest implementation, the hybrid approach constructs a matched sample based on these characteristics, followed by applying DiD on the matched sample to difference out any remaining unobserved time-invariant factors. Thus, matching mitigates concerns that differential trends, rather than the policy intervention, drive observed differences in outcomes. Since the 2v3 and 3v4 samples have different numbers of treatment and control variables, I test the effectiveness of different matching procedures for each.

4.3.1 2v3 Sample

In the 2v3 sample, there are 1,930 control units and 5,010 treated units. This distribution reflects the relative rarity of the 2-year math requirement. This imbalance poses challenges for 1:1 nearest neighbor matching, a simple and easily interpretable matching method that pairs each treated unit with the most similar control unit based on covariates. While computationally efficient and easy to interpret, 1:1 matching would require discarding a substantial number of treated observations in this case, potentially reducing statistical power and representativeness. To address this, I also implement full matching, which allows for flexible weighting of multiple control and treated units within matched strata. I then compare the covariate balance achieved by each method to evaluate their relative performance.

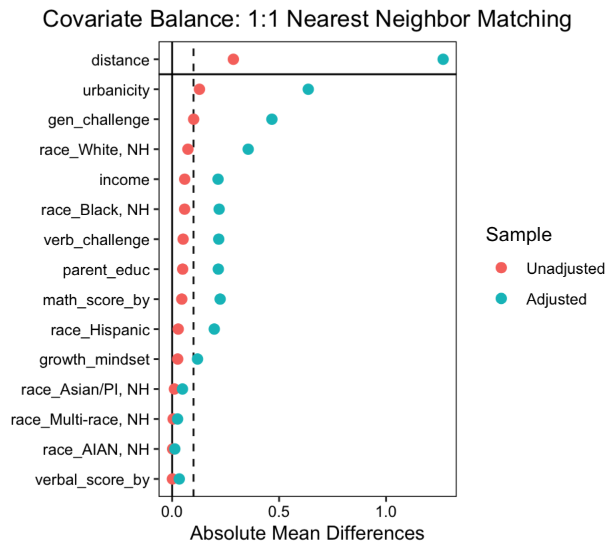


Figure 4.1: Covariate Balance (2v3 Sample): 1:1 Nearest Neighbor Matching

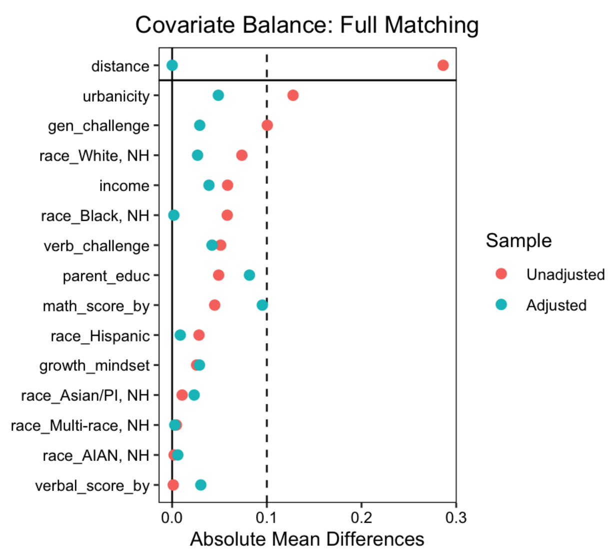


Figure 4.2: Covariate Balance (2v3 Sample): Full Matching

To evaluate the performance of different matching methods, I compare covariate balance between treated and control units after adjustment using absolute mean differences (AMDs). AMDs measure the absolute difference in means for each covariate between treatment groups, standardized by the pooled standard deviation. Smaller values indicate better balance, and a common rule of thumb is that covariates with AMDs below 0.1 are considered adequately balanced.

Figures 4.1 and 4.2 show AMDs before and after matching for the 2v3 sample, using 1:1 nearest neighbor matching and full matching, respectively. In the case of 1:1 matching (Figure 4.1), several covariates exhibit worse balance after adjustment. This likely occured because the method discards a large number of treated units (due to the smaller control pool), reducing representativeness and increasing sampling variance, especially when matches are suboptimal. As a result, only 4 out of 14 covariates achieve balance after matching, with 10 remaining above the 0.1 threshold.

In contrast, full matching (Figure 4.2) yields a much better covariate balance across the board. All 14 covariates fall below the 0.1 threshold after adjustment, indicating strong balance. Full matching uses all available data and flexibly forms matched strata of varying sizes, assigning weights accordingly. This enables tighter matches without discarding units, leading to more robust and representative comparisons between the treated and control groups. Given the substantial improvement in covariate balance, I proceed with full matching for the 2v3 analysis.

Table 4.1: Comparison of Covariate Balance (2v3 Sample): 1:1 Nearest Neighbor vs Full Matching

Variable	**1:1 Matching**		**Full Matching**	
	Diff (Adj.)	Balance Status	Diff (Adj.)	Balance Status
White	-0.3554	Not Balanced (>0.1)	-0.0268	Balanced (<0.1)
Black	0.2197	Not Balanced (>0.1)	-0.0018	Balanced (<0.1)
Hispanic	0.1969	Not Balanced (>0.1)	0.0086	Balanced (<0.1)
Asian/Pacific Islander	-0.0482	Balanced (<0.1)	0.0233	Balanced (<0.1)
American Indian/Alaska Native	0.0124	Balanced (<0.1)	-0.0060	Balanced (<0.1)
Multiracial	-0.0254	Balanced (<0.1)	0.0028	Balanced (<0.1)
Parent Education	0.2158	Not Balanced (>0.1)	0.0816	Balanced (<0.1)
Household Income	0.2146	Not Balanced (>0.1)	0.0389	Balanced (<0.1)
Urbanicity	-0.6363	Not Balanced (>0.1)	0.0488	Balanced (<0.1)
General Challenge	0.4663	Not Balanced (>0.1)	0.0291	Balanced (<0.1)
Verbal SE	0.2177	Not Balanced (>0.1)	-0.0421	Balanced (<0.1)
Growth Mindset	-0.1186	Not Balanced (>0.1)	-0.0289	Balanced (<0.1)
Math Score (10th Grade)	-0.2243	Not Balanced (>0.1)	0.0952	Balanced (<0.1)
Verbal Score (10th Grade)	-0.0333	Balanced (<0.1)	0.0303	Balanced (<0.1)

Balance Tally	**1:1 Matching**	**Full Matching**
Balanced (<0.1)	4	15
Not Balanced (>0.1)	10	0

Sample Sizes	**1:1 Matching**	**Full Matching**
Control Units (All)	1930	1930
Treated Units (All)	5010	5010
Matched Control Units	1930	368.13 (ESS)
Matched Treated Units	1930	5010

Note: Under full matching, all treated and control units are retained. However, units are assigned weights based on the quality of the match, such that some poorly matched controls receive very small weights. The "Matched (ESS)" figure for controls reflects the Effective Sample Size (ESS), which adjusts for these weights. An ESS of 368.13 indicates that, although 1,930 control units are present, their combined weighted contribution is equivalent to approximately 368 fully weighted controls. ESS is calculated as $(\sum w_i)^2/\sum w_i^2$, *where* w_i *are the weights assigned to each unit.*

4.3.2 3v4 Sample

For the 3v4 sample, there are 5,010 control units and 1,215 treated units, resulting in a larger control pool relative to the treated group. To take advantage of the available data and improve efficiency, I implement 1:3 nearest neighbor matching, allowing each treated unit to be matched to up to three control units with the closest propensity scores. This approach increases the number of matched control observations while retaining all treated units. Covariate balance after matching is strong: all 14 covariates have absolute mean differences below the 0.1 threshold, as shown in Figure 4.3 and detailed in Table 4.2. These results indicate that 1:3 nearest neighbor matching achieves adequate balance and preserves a large portion of the sample, making it suitable for analysis in this setting.

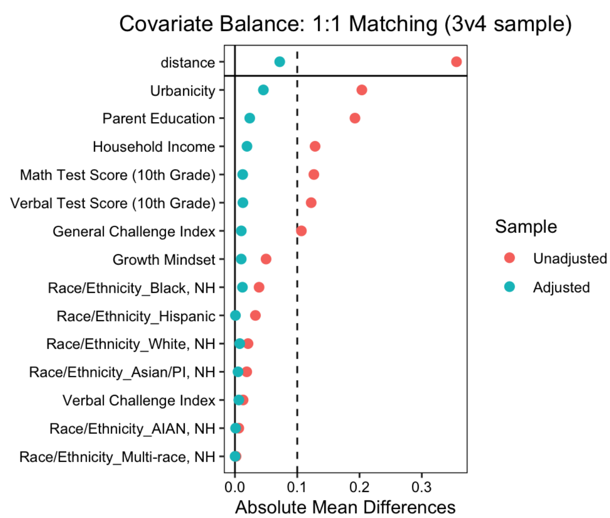


Figure 4.3: Covariate Balance (3v4 Sample): 1:3 Nearest Neighbor Matching

Table 4.2: Balance Measures for 1:3 Nearest Neighbor Matching (3v4 Sample)

Variable	**Diff (Adj.)**	**Balance Status**
Distance	0.0719	Balanced (<0.1)
White	-0.0077	Balanced (<0.1)
Black	0.0121	Balanced (<0.1)
Hispanic	-0.0008	Balanced (<0.1)
Asian/Pacific Islander	-0.0049	Balanced (<0.1)
American Indian/Alaska Native	0.0011	Balanced (<0.1)
Multiracial	0.0003	Balanced (<0.1)
Parent Education	0.0238	Balanced (<0.1)
Household Income	0.0193	Balanced (<0.1)
Urbanicity	-0.0456	Balanced (<0.1)
General Challenge	0.0102	Balanced (<0.1)
Verbal SE	-0.0062	Balanced (<0.1)
Growth Mindset	0.0101	Balanced (<0.1)
Math Score (10th Grade)	0.0125	Balanced (<0.1)
Verbal Score (10th Grade)	0.0128	Balanced (<0.1)

Sample Sizes	**Control**	**Treated**
All	5010	1215
Matched	3645	1215
Unmatched	1365	0

CHAPTER 5
RESULTS

5.1 2 to 3 Year Sample

In the comparison between states requiring two versus three years of high school mathematics, I find no significant treatment effect on students' math self-efficacy overall. Across all specifications, the coefficient on the interaction term Treated × Post remains small and statistically insignificant, suggesting that increasing math requirements from two to three years did not have a discernible causal impact on students' average self-efficacy.

However, gender disparities are pronounced throughout the models. Across specifications, the coefficient on Female is consistently negative and statistically significant, highlighting a persistent self-efficacy gap disadvantaging female students. Even after controlling for demographic and psychological covariates, and applying both lagged outcomes and propensity score matching, female students reported substantially lower math self-efficacy than their male peers. This aligns with prior literature documenting gendered patterns of internalized doubt in mathematical ability, and underscores the need to interpret math policy reforms through a lens that accounts for how such reforms interact with structural inequities.

In the controlled model, psychological traits such as general challenge index, verbal self-efficacy, and growth mindset are significantly associated with higher math self-efficacy. However, once I control for baseline (10th grade) math self-efficacy in the lagged model, these associations largely disappear. This suggests that the predictive power of these traits was confounded by pre-existing differences in math self-efficacy. After matching, growth mindset in particular loses significance, implying that it reflected selection bias rather than a causal mechanism.

Given these persistent gender gaps and confounded relationships between predictors and outcomes, I turn to a triple interaction model to explore whether treatment effects varied

Table 5.1: Regression Results: 2v3 Sample

	Naïve (1)	+Controls (2)	+Lagged (3)	Matching (4)	Matching+Lagged (5)
Treated	-0.006 (0.028)	-0.002 (0.021)	0.006 (0.008)	-0.007 (0.058)	-0.002 (0.017)
Post	0.011 (0.032)	0.008 (0.029)	0.011 (0.023)	0.056 (0.072)	0.079 (0.055)
Treated × Post	0.002 (0.038)	-0.009 (0.033)	-0.008 (0.027)	-0.067 (0.076)	-0.080 (0.058)
Female		-0.183*** (0.015)	-0.021$^+$ (0.012)	-0.275*** (0.030)	-0.079*** (0.022)
Black		0.068* (0.028)	0.019 (0.022)	0.076 (0.050)	0.046 (0.037)
Hispanic		0.053* (0.025)	0.015 (0.018)	0.000 (0.047)	-0.028 (0.034)
Asian/PI		-0.033 (0.028)	-0.017 (0.022)	0.046 (0.053)	0.060* (0.031)
AIAN		-0.049 (0.076)	-0.052 (0.061)	0.112 (0.123)	0.011 (0.087)
Multi-race		-0.004 (0.038)	0.003 (0.026)	-0.175$^+$ (0.105)	0.052 (0.099)
Parent Education		0.009* (0.004)	0.003 (0.003)	-0.003 (0.008)	-0.003 (0.006)
Income		-0.002 (0.004)	-0.003 (0.003)	-0.004 (0.007)	0.004 (0.005)
Urbanicity		-0.011 (0.012)	-0.004 (0.009)	-0.018 (0.022)	-0.019 (0.016)
General Challenge		0.358*** (0.015)	0.014 (0.012)	0.015 (0.025)	-0.028 (0.017)
Verbal SE		0.072*** (0.013)	-0.010 (0.009)	-0.030 (0.022)	0.006 (0.014)
Growth Mindset		-0.170*** (0.012)	-0.032** (0.010)	-0.002 (0.023)	-0.022 (0.018)
Math Score (10th Grade)		0.029*** (0.001)	0.009*** (0.001)	0.004* (0.002)	-0.000 (0.001)
Verbal Score (10th Grade)		-0.014*** (0.001)	-0.003** (0.001)	-0.000 (0.002)	0.001 (0.001)
Math Self-Efficacy (10th Grade)			0.698*** (0.010)		0.728*** (0.015)
Constant	2.492*** (0.024)	1.048*** (0.065)	0.543*** (0.050)	2.628*** (0.127)	0.801*** (0.100)
Observations	15625	13320	13317	6741	5914
R^2	0.000	0.301	0.586	0.034	0.570

Note: Robust standard errors in parentheses. $^+p < 0.1$, $*p < 0.05$, $**p < 0.01$, $***p < 0.001$

Table 5.2: Triple Interaction Regression Results: 2v3 Sample

	Naïve Tri (1)	+Controls Tri (2)	+Lagged Tri (3)	Matching Tri (4)	Matching+Lagged Tri (5)
Treated	-0.067$^+$ (0.039)	-0.046 (0.029)	-0.012 (0.010)	-0.109 (0.092)	-0.028 (0.027)
Post	-0.038 (0.047)	-0.048 (0.041)	-0.045 (0.032)	-0.000 (0.115)	0.109 (0.086)
Female	-0.351*** (0.047)	-0.278*** (0.036)	-0.083*** (0.013)	-0.435*** (0.110)	-0.115*** (0.032)
Treated × Post	0.004 (0.054)	0.011 (0.048)	0.012 (0.038)	-0.046 (0.120)	-0.133 (0.090)
Treated × Female	0.116* (0.054)	0.084* (0.042)	0.035* (0.015)	0.185 (0.116)	0.049 (0.033)
Post × Female	0.081 (0.064)	0.106$^+$ (0.057)	0.107* (0.045)	0.094 (0.143)	-0.056 (0.111)
Treated × Post × Female	0.001 (0.074)	-0.039 (0.067)	-0.039 (0.053)	-0.026 (0.152)	0.099 (0.117)
Black		0.068* (0.028)	0.018 (0.022)	0.075 (0.050)	0.046 (0.037)
Hispanic		0.053* (0.025)	0.015 (0.018)	-0.011 (0.047)	-0.035 (0.034)
Asian/PI		-0.032 (0.028)	-0.016 (0.022)	0.045 (0.052)	0.060* (0.030)
AIAN		-0.048 (0.077)	-0.052 (0.061)	0.118 (0.122)	0.017 (0.088)
Multi-race		-0.004 (0.038)	0.003 (0.026)	-0.180$^+$ (0.106)	0.049 (0.096)
Parent Education		0.009* (0.004)	0.003 (0.003)	-0.004 (0.008)	-0.004 (0.006)
Income		-0.002 (0.004)	-0.003 (0.003)	-0.005 (0.007)	0.003 (0.005)
Urbanicity		-0.010 (0.012)	-0.004 (0.009)	-0.016 (0.022)	-0.018 (0.016)
General Challenge		0.358*** (0.015)	0.014 (0.012)	0.015 (0.024)	-0.028 (0.017)
Verbal Self-Efficacy		0.072*** (0.013)	-0.010 (0.009)	-0.030 (0.022)	0.006 (0.014)
Growth Mindset		-0.170*** (0.012)	-0.032** (0.010)	-0.001 (0.023)	-0.022 (0.018)
Math Score (10th Grade)		0.029*** (0.001)	0.009*** (0.001)	0.004* (0.002)	-0.000 (0.001)
Verbal Score (10th Grade)		-0.014*** (0.001)	-0.003** (0.001)	-0.000 (0.002)	0.001 (0.001)
Math Self-Efficacy (10th Grade)			0.698*** (0.010)		0.727*** (0.015)
Constant	2.676*** (0.034)	1.099*** (0.066)	0.576*** (0.050)	2.719*** (0.145)	0.826*** (0.101)
Observations	15614	13320	13317	6741	5914
R^2	0.020	0.301	0.586	0.037	0.571

Note: Robust standard errors in parentheses. $^+p < 0.1$, $*p < 0.05$, $**p < 0.01$, $***p < 0.001$

by gender. Although the three-way interaction term Treated × Post × Female is not statistically significant, two lower-order interactions—Treated × Female and Post × Female—are significant in early specifications. These results suggest that female students in treated states already exhibited lower self-efficacy in control states compared to males, a gap that was wider than in control states. They also suggest that female students' math self-efficacy evolved increased less over time, regardless of treatment. This gendered time trend may reflect cultural or developmental processes that influence self-perception during adolescence and may not be addressed by increasing course requirements alone.

5.2 3 to 4 Year Sample

In the comparison between states requiring three versus four years of high school mathematics, I find evidence of a positive treatment effect on students' math self-efficacy. The interaction term Treated × Post becomes marginally significant with the inclusion of the lagged variable.

Female students again report significantly lower levels of math self-efficacy, though the magnitude of this gap decreases with the inclusion of lagged controls. In contrast to the 2v3 sample, where most psychological covariates lost significance after controlling for baseline self-efficacy, both math and verbal test scores remain significant predictors in the 3v4 models. This suggests that the predictive relationship between academic performance and math self-efficacy may be more robust in this sample, or that the baseline confounding was less severe.

Psychological traits such as general challenge, verbal self-efficacy, and growth mindset are significant in the controlled models but lose explanatory power once prior self-efficacy is added. This again highlights the importance of accounting for baseline beliefs in identifying the role of psychological traits, as their apparent effects may largely reflect earlier differences in self-perception.

Turning to the triple interaction models, I find that the treatment effect becomes stronger

Table 5.3: Regression Results: 3v4 Sample

	Naïve (1)	+Controls (2)	+Lagged (3)	Matching (4)	Matching+Lagged (5)
Treated	0.048^{+} (0.028)	0.042^{+} (0.022)	0.013^{+} (0.008)	0.042^{+} (0.022)	0.013^{+} (0.008)
Post	-0.012 (0.020)	-0.016 (0.017)	-0.014 (0.013)	-0.016 (0.017)	-0.014 (0.013)
Treated × Post	0.053 (0.040)	0.053 (0.034)	0.052^{+} (0.027)	0.053 (0.034)	0.052^{+} (0.027)
Female		-0.170*** (0.015)	-0.029* (0.012)	-0.170*** (0.015)	-0.029* (0.012)
Black		0.083*** (0.024)	0.045* (0.019)	0.083*** (0.024)	0.045* (0.019)
Hispanic		0.026 (0.026)	0.034^{+} (0.019)	0.026 (0.026)	0.034^{+} (0.019)
Asian/PI		-0.061* (0.028)	-0.018 (0.021)	-0.061* (0.028)	-0.018 (0.021)
AIAN		0.394^{+} (0.226)	-0.071 (0.216)	0.394^{+} (0.226)	-0.071 (0.216)
Multi-race		-0.017 (0.040)	0.010 (0.029)	-0.017 (0.040)	0.010 (0.029)
Parent Education		0.001 (0.004)	0.004 (0.003)	0.001 (0.004)	0.004 (0.003)
Income		-0.000 (0.004)	-0.001 (0.003)	-0.000 (0.004)	-0.001 (0.003)
Urbanicity		-0.008 (0.011)	-0.004 (0.009)	-0.008 (0.011)	-0.004 (0.009)
General Challenge		0.350*** (0.014)	0.014 (0.011)	0.350*** (0.014)	0.014 (0.011)
Verbal Self-Efficacy		0.090*** (0.012)	0.004 (0.009)	0.090*** (0.012)	0.004 (0.009)
Growth Mindset		-0.142*** (0.012)	-0.011 (0.009)	-0.142*** (0.012)	-0.011 (0.009)
Math Score (10th Grade)		0.030*** (0.001)	0.009*** (0.001)	0.030*** (0.001)	0.009*** (0.001)
Verbal Score (10th Grade)		-0.015*** (0.001)	-0.002* (0.001)	-0.015*** (0.001)	-0.002* (0.001)
Math Self-Efficacy (10th Grade)			0.691*** (0.010)		0.691*** (0.010)
Constant	2.589*** (0.014)	0.979*** (0.063)	0.480*** (0.048)	0.979*** (0.063)	0.480*** (0.048)
Observations	9389	9389	9386	9389	9386
R^2	0.002	0.281	0.575	0.281	0.575

Note: Robust standard errors in parentheses. $^{+}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

at a level of $p < 0.05$ when gender interaction terms are included. This suggests that the initial DiD models were masking gender-specific dynamics, and that gender acted as an effect modifier. In particular, the coefficient on Post × Female is positive and significant across all specifications, indicating that female students' math self-efficacy improved more over time than that of male students, regardless of treatment. This positive time trend may have contributed to the observed treatment effect and confounded earlier estimates.

The three-way interaction term Treated × Post × Female remains statistically insignificant, implying that the treatment effect itself was not significantly different for girls and boys. However, the significance of the Post × Female term signals important gendered changes over time, reinforcing the value of modeling interaction effects. These results support the decision to include gender interactions in the analysis, as they reveal patterns that would otherwise remain obscured.

Table 5.4: Triple Interaction Regression Results: 3v4 Sample

	Naïve Tri (1)	+Controls Tri (2)	+Lagged Tri (3)	Matching Tri (4)	Matching+Lagged Tri (5)
Treated	0.030 (0.041)	0.043 (0.031)	0.011 (0.011)	0.043 (0.031)	0.011 (0.011)
Post	-0.055$^+$ (0.029)	-0.059* (0.025)	-0.057** (0.020)	-0.059* (0.025)	-0.057** (0.020)
Female	-0.262*** (0.028)	-0.200*** (0.022)	-0.061*** (0.009)	-0.200*** (0.022)	-0.061*** (0.009)
Treated × Post	0.087 (0.058)	0.089$^+$ (0.049)	0.086* (0.038)	0.089$^+$ (0.049)	0.086* (0.038)
Treated × Female	0.034 (0.056)	-0.002 (0.045)	0.004 (0.016)	-0.002 (0.045)	0.004 (0.016)
Post × Female	0.079* (0.040)	0.081* (0.034)	0.080** (0.027)	0.081* (0.034)	0.080** (0.027)
Treated × Post × Female	-0.063 (0.079)	-0.066 (0.068)	-0.064 (0.054)	-0.066 (0.068)	-0.064 (0.054)
Black		0.083*** (0.024)	0.044* (0.019)	0.083*** (0.024)	0.044* (0.019)
Hispanic		0.026 (0.026)	0.034$^+$ (0.019)	0.026 (0.026)	0.034$^+$ (0.019)
Asian/PI		-0.061* (0.028)	-0.017 (0.021)	-0.061* (0.028)	-0.017 (0.021)
AIAN		0.400$^+$ (0.230)	-0.067 (0.221)	0.400$^+$ (0.230)	-0.067 (0.221)
Multi-race		-0.016 (0.040)	0.010 (0.029)	-0.016 (0.040)	0.010 (0.029)
Parent Education		0.001 (0.004)	0.004 (0.003)	0.001 (0.004)	0.004 (0.003)
Income		-0.000 (0.004)	-0.002 (0.003)	-0.000 (0.004)	-0.002 (0.003)
Urbanicity		-0.008 (0.011)	-0.004 (0.009)	-0.008 (0.011)	-0.004 (0.009)
General Challenge		0.350*** (0.014)	0.013 (0.011)	0.350*** (0.014)	0.013 (0.011)
Verbal Self-Efficacy		0.090*** (0.012)	0.004 (0.009)	0.090*** (0.012)	0.004 (0.009)
Growth Mindset		-0.142*** (0.012)	-0.012 (0.009)	-0.142*** (0.012)	-0.012 (0.009)
Math Score (10th Grade)		0.030*** (0.001)	0.009*** (0.001)	0.030*** (0.001)	0.009*** (0.001)
Verbal Score (10th Grade)		-0.015*** (0.001)	-0.002* (0.001)	-0.015*** (0.001)	-0.002* (0.001)
Math Self-Efficacy (10th Grade)			0.691*** (0.010)		0.691*** (0.010)
Constant	2.729*** (0.020)	0.997*** (0.063)	0.498*** (0.048)	0.997*** (0.063)	0.498*** (0.048)
Observations	9389	9389	9386	9389	9386
R^2	0.020	0.282	0.576	0.282	0.576

Note: Robust standard errors in parentheses. $^+p < 0.1$, $*p < 0.05$, $**p < 0.01$, $***p < 0.001$

5.3 Exploring Heterogeneous Treatment Effects of Race, Gender, and Baseline Self-efficacy

To further investigate the heterogeneous effects of math graduation requirements on students' math self-efficacy, I included interaction terms with race, gender, and baseline math self-efficacy. This decision was informed by several patterns in earlier models. In the 2v3 sample, the addition of a triple interaction between treatment, post, and gender revealed that the treatment effect became larger and statistically significant when gender was included, suggesting that gender differences were masking the true effect in simpler DiD models. Additionally, in both the 2v3 and 3v4 samples, the coefficient on female was consistently negative and highly significant, indicating persistent gender gaps in self-efficacy regardless of treatment status. These patterns motivated the inclusion of gender interactions.

Race-based heterogeneity also appeared in earlier models. In the 3v4 sample, several racial groups exhibited significantly different levels of self-efficacy. For instance, Black and Hispanic students had significantly higher self-efficacy compared to the reference group, even after controlling for covariates. Moreover, in the 3v4 triple interaction model, the treatment effect for Hispanic and Asian/PI students was significantly smaller than for White students, as shown by the negative and marginally significant coefficients on Treated $\times$ Hispanic and Treated $\times$ Asian/PI. In the 2v3 sample, the triple interaction Treated $\times$ Post $\times$ Black was significantly negative, indicating that Black students experienced a significantly smaller treatment effect than White students in response to the increased math requirement.

These subgroup differences in both baseline levels and treatment responsiveness raised concerns that pooled models might obscure important variation. Therefore, I introduced interaction terms to assess whether specific race or gender groups experienced systematically different effects of treatment and whether the predictive relationship between grade 10 math self-efficacy and Grade 12 outcomes varied by subgroup. This was especially important given the strong predictive power of baseline self-efficacy, and evidence from the literature suggesting that minoritized groups face stereotype threat and other structural barriers that may moderate how they internalize or respond to policy changes.

I first explored race-based heterogeneity by including triple interactions between treatment, post-period, and race. In the 2v3 sample, I found that Black students experienced a significantly negative treatment effect in the post period, with a coefficient of -0.361 ($p<0.05$) for Treated × Post × Black. No other racial group displayed significant triple interaction effects, though the large negative coefficient for multi-racial students (-0.768) approached significance. These results suggest that the policy may have particularly undermined self-efficacy gains among Black students, potentially exacerbating existing disparities.

When interacting race with gender, I observed marginally significant positive effects for Hispanic × Female ($p<0.1$) and AIAN × Female ($p<0.1$) in the 2v3 sample, which may reflect group-specific resilience or unobserved protective factors. However, these results did not generalize to the 3v4 sample, where none of the race × gender terms were significant. The most notable 3v4 finding came from the triple interaction for multi-racial students, where Treated × Post × Multi-race was positive and marginally significant (0.275, $p<0.1$), contrasting with the 2v3 result. This could suggest a shift in which students are most affected by policy changes depending on where in the distribution the graduation threshold is being raised.

To assess whether baseline math self-efficacy differentially predicts outcomes across groups, I also interacted Grade 10 self-efficacy with race and gender. These interactions test whether the effect of early self-perceptions depends on demographic background, potentially explaining why some students improve while others stagnate. However, in both 2v3 and 3v4 samples, none of the interaction terms between baseline self-efficacy and race or gender were statistically significant. The coefficients were generally small in magnitude, and the model fit did not improve meaningfully with these additions. This suggests that while gender and race may shape the overall levels and trajectories of self-efficacy, they do not significantly

Table 5.5: Race-Based Heterogeneity in 12th Grade Math Self-Efficacy (2v3 Sample)

	Treated × Post × Race (1)	Race × Gender (2)
Treated	-0.001 (0.019)	-0.004 (0.017)
Post	0.015 (0.068)	0.079 (0.055)
Treated × Post	-0.014 (0.072)	-0.081 (0.058)
Female	-0.080*** (0.021)	-0.103*** (0.029)
Black	0.040 (0.049)	0.039 (0.051)
Hispanic	-0.041 (0.044)	-0.082 (0.051)
Asian/PI	0.033 (0.066)	0.034 (0.042)
AIAN	0.018 (0.035)	-0.096 (0.100)
Multi-race	-0.169 (0.107)	0.042 (0.159)
Treated × Black	-0.037 (0.051)	
Treated × Hispanic	0.061 (0.046)	
Treated × Asian/PI	-0.038 (0.068)	
Treated × AIAN	-0.022 (0.059)	
Treated × Multi-race	0.133 (0.109)	
Post × Black	0.304^{+} (0.157)	
Post × Hispanic	-0.047 (0.150)	
Post × Asian/PI	0.146 (0.135)	
Post × AIAN	0.165 (0.413)	
Post × Multi-race	0.805 (0.506)	
Treated × Post × Black	-0.361* (0.167)	
Treated × Post × Hispanic	-0.017 (0.161)	
Treated × Post × Asian/PI	-0.057 (0.147)	
Treated × Post × AIAN	-0.208 (0.447)	
Treated × Post × Multi-race	-0.768 (0.515)	
Black × Female		0.014 (0.068)
Hispanic × Female		0.110^{+} (0.063)
Asian/PI × Female		0.052 (0.059)
AIAN × Female		0.318^{+} (0.167)
Multi-race × Female		0.017 (0.167)
Parent Education	-0.005 (0.005)	-0.004 (0.006)
Income	0.005 (0.005)	0.003 (0.005)
Urbanicity	-0.023 (0.016)	-0.018 (0.016)
General Challenge	-0.027 (0.017)	-0.028 (0.017)
Verbal Self-Efficacy	0.008 (0.014)	0.008 (0.014)
Growth Mindset	-0.023 (0.018)	-0.020 (0.018)
Math Score (10th Grade)	-0.000 (0.001)	-0.000 (0.001)
Verbal Score (10th Grade)	0.001 (0.001)	0.001 (0.001)
Math Self-Efficacy (10th Grade)	0.729*** (0.013)	0.728*** (0.014)
Constant	0.809*** (0.089)	0.808*** (0.097)
Observations	5914	5914
R^2	0.576	0.570
Adj. R^2	0.573	0.569
AIC	10314.6	10321.1
BIC	45768.2	47564.7
Log Likelihood	-22732.1	-23673.8
RMSE	0.55	0.55

Note: Robust standard errors in parentheses. $^{+}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Table 5.6: Interaction of 10th Grade Math Self-Efficacy with Race and Gender (2v3 Sample)

	M10 × Race (1)	M10 × Gender (2)
Treated	0.001 (0.017)	0.001 (0.017)
Post	0.079 (0.055)	0.079 (0.055)
Treated × Post	-0.080 (0.057)	-0.080 (0.058)
Female	-0.077*** (0.022)	-0.089 (0.082)
10th Grade Self-Efficacy	0.733*** (0.016)	0.726*** (0.025)
Black	0.030 (0.114)	
Hispanic	-0.017 (0.110)	
Asian/PI	0.053 (0.096)	
AIAN	0.004 (0.229)	
Multi-race	0.347 (0.446)	
10th Grade SE × Black	0.007 (0.040)	
10th Grade SE × Hispanic	-0.004 (0.045)	
10th Grade SE × Asian/PI	0.003 (0.036)	
10th Grade SE × AIAN	0.004 (0.086)	
10th Grade SE × Multi-race	-0.129 (0.158)	
10th Grade SE × Female		0.003 (0.030)
Parent Education	-0.004 (0.006)	-0.001 (0.006)
Income	0.004 (0.005)	0.001 (0.005)
Urbanicity	-0.021 (0.016)	-0.022 (0.015)
General Challenge	-0.027 (0.017)	-0.026 (0.018)
Verbal Self-Efficacy	0.007 (0.014)	0.004 (0.014)
Growth Mindset	-0.022 (0.018)	-0.024 (0.018)
Math Score (10th Grade)	-0.000 (0.001)	0.000 (0.001)
Verbal Score (10th Grade)	0.001 (0.001)	0.001 (0.001)
Constant	0.779*** (0.095)	0.835*** (0.118)
Observations	5914	5914
R^2	0.570	0.569
Adj. R^2	0.569	0.568
AIC	10342.0	10299.4
BIC	47256.8	47760.1
Log Likelihood	-23519.8	-23810.6
RMSE	0.55	0.55

Note: Robust standard errors in parentheses. $^{+}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

moderate the predictive power of early self-efficacy itself.

Altogether, these findings indicate that the treatment effect is not uniform across student subgroups. I found evidence that gender and race, particularly in combination, may influence how students respond to higher math requirements. The negative impact for Black students in the 2v3 sample is especially concerning, reinforcing the need to analyze policy through an intersectional lens. These results support the inclusion of interaction terms to ensure that heterogeneous responses are not masked by average treatment effects.

Table 5.7: Race-Based Heterogeneity in Grade-12 Math Self-Efficacy (3v4 Sample)

	Treated × Post × Race (1)	Race × Gender (2)
Intercept	0.476*** (0.048)	0.481*** (0.048)
Treated	0.022* (0.010)	0.013^{+} (0.008)
Post	−0.004 (0.017)	−0.014 (0.013)
Black	0.048*** (0.014)	0.037 (0.028)
Hispanic	0.037* (0.015)	0.045^{+} (0.027)
Asian/PI	0.017 (0.015)	0.001 (0.030)
AIAN	0.123 (0.196)	0.350 (0.359)
Multi-race	0.009 (0.020)	−0.017 (0.038)
Female	−0.029* (0.012)	-0.027^{+} (0.015)
Parent Education	0.004 (0.003)	0.004 (0.003)
Income	−0.001 (0.003)	−0.002 (0.003)
Urbanicity	−0.003 (0.009)	−0.004 (0.009)
General Challenge	0.013 (0.011)	0.013 (0.011)
Verbal Self-Efficacy	0.004 (0.009)	0.004 (0.009)
Growth Mindset	−0.011 (0.009)	−0.011 (0.009)
Math Score (10th Grade)	0.009*** (0.001)	0.009*** (0.001)
Verbal Score (10th Grade)	−0.002* (0.001)	−0.002* (0.001)
Math SE (10th Grade)	0.691*** (0.010)	0.691*** (0.010)
Treated × Post	0.026 (0.034)	0.052^{+} (0.027)
Race interactions		
Treated × Black	0.015 (0.024)	
Treated × Hispanic	-0.049^{+} (0.027)	
Treated × Asian/PI	-0.051^{+} (0.029)	
Treated × AIAN	0.040 (0.214)	
Treated × Multi-race	−0.069 (0.043)	
Post × Black	−0.028 (0.041)	
Post × Hispanic	−0.009 (0.044)	
Post × Asian/PI	−0.044 (0.047)	
Post × AIAN	−0.329 (0.772)	
Post × Multi-race	−0.032 (0.069)	
Treated × Post × Black	0.044 (0.080)	
Treated × Post × Hispanic	0.110 (0.087)	
Treated × Post × Asian/PI	−0.014 (0.101)	
Treated × Post × AIAN	−0.193 (0.803)	
Treated × Post × Multi-race	0.275^{+} (0.143)	
Gender–Race interactions		
Black × Female		0.012 (0.035)
Hispanic × Female		−0.021 (0.037)
Asian/PI × Female		−0.037 (0.040)
AIAN × Female		−0.562 (0.430)
Multi-race × Female		0.053 (0.059)
Observations	9 386	9 386
R^2	0.576	0.576
Adjusted R^2	0.575	0.575
AIC	15 317.7	15 305.2
BIC	26 450.9	26 322.8
Log–Likelihood	−13 065.389	−13 047.073
F Statistic	365.192	507.398
RMSE	0.55	0.55

Note: Robust standard errors in parentheses. $^{+}p < 0.10$, $*p < 0.05$, $**p < 0.01$, $***p < 0.001$

Table 5.8: Interaction of Grade-10 Math Self-Efficacy with Race and Gender (3v4 Sample)

	Grade-10 SE × Race (1)	Grade-10 SE × Gender (2)
Intercept	0.475*** (0.050)	0.484*** (0.051)
Treated	0.013^{+} (0.008)	0.013^{+} (0.008)
Post	−0.014 (0.013)	−0.014 (0.013)
Grade-10 SE	0.693*** (0.011)	0.689*** (0.013)
Black	0.101 (0.062)	0.045* (0.019)
Hispanic	−0.018 (0.064)	0.034^{+} (0.019)
Asian/PI	−0.004 (0.075)	−0.018 (0.021)
AIAN	1.003 (0.938)	−0.072 (0.216)
Multi-race	0.050 (0.101)	0.010 (0.029)
Female	−0.029* (0.012)	−0.036 (0.040)
Parent Education	0.004 (0.003)	0.004 (0.003)
Income	−0.002 (0.003)	−0.001 (0.003)
Urbanicity	−0.004 (0.009)	−0.004 (0.009)
General Challenge	0.014 (0.011)	0.014 (0.011)
Verbal Self-Efficacy	0.004 (0.009)	0.004 (0.009)
Growth Mindset	−0.011 (0.009)	−0.011 (0.009)
Math Score (10th Grade)	0.009*** (0.001)	0.009*** (0.001)
Verbal Score (10th Grade)	−0.002* (0.001)	−0.002* (0.001)
Treated × Post	0.052^{+} (0.027)	0.052^{+} (0.027)
Grade-10 SE × Race interactions		
SE × Black	−0.022 (0.024)	
SE × Hispanic	0.020 (0.023)	
SE × Asian/PI	−0.005 (0.027)	
SE × AIAN	−0.349 (0.320)	
SE × Multi-race	−0.016 (0.037)	
Grade-10 SE × Gender interaction		
SE × Female		0.003 (0.015)
Observations	9 386	9 386
R^2	0.576	0.575
Adjusted R^2	0.575	0.575
AIC	15 307.6	15 302.2
BIC	26 323.3	26 271.6
Log–Likelihood	−13 047.329	−13 039.745
F Statistic	510.154	612.641
RMSE	0.55	0.55

Note: Robust standard errors in parentheses. $^{+}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

CHAPTER 6

CONCLUSION

6.1 Implications

The results of this study carry several implications for education policy, particularly regarding the design of high school math graduation requirements and their impact on student confidence in math.

The comparison between states requiring two versus three years of math showed no significant treatment effects, suggesting that this policy shift may not meaningfully influence students' math self-efficacy. One likely explanation is that the additional year of coursework, usually Geometry or Algebra II, may not represent a sufficiently challenging or novel curricular change to generate meaningful gains in confidence. This interpretation is supported by Table 3.6, which shows relatively modest changes in course-taking between the 2- and 3-year requirement groups, especially compared to the more pronounced shifts observed in the 3v4 comparison. As well, I reasoned that many students in the 2-year states were already voluntarily enrolling in three or more years of math, reducing the policy's marginal effect.

Moreover, Figure 3.7 suggests that courses such as geometry can have mixed or even negative effects on self-efficacy, particularly for female students in the deciles of middle ability. This raises the possibility that the students most likely to be affected by the 2v3 requirement (i.e., those nudged from Algebra I into Geometry) are not necessarily experiencing confidence gains and may, in fact, be confronting new challenges without adequate support. In contrast, students at the top of the distribution, many of whom would have taken advanced math regardless of the policy, may already be on trajectories where the coursework reinforces confidence. Thus, the absence of an average treatment effect may reflect both limited compliance and heterogeneous effects across the ability distribution, where true compliers are more likely to experience flat or negative changes in self-efficacy. It is also important to note that very few states maintained a two-year requirement during the period of observation, so the 2v3 comparison may suffer from limited sample size or sample selection bias, making those states potentially unrepresentative.

Although no average treatment effect was detected in the 2v3 sample, I found consistent and significant gender gaps in math self-efficacy, with female students reporting lower confidence than their male peers across all models. Triple interaction models further revealed that these gender gaps were partially explained by differences in baseline levels (Treated $\times$ Female) and time trends (Post $\times$ Female), even in the absence of a gendered treatment effect. These findings suggest that female students' math confidence may evolve differently over time due to factors unrelated to policy changes, such as stereotype threat or broader cultural narratives about gender and math ability in these states. The persistence of gender disparities, even after accounting for a wide array of psychological and academic controls, underscores the importance of complementary interventions, such as targeted mentorship or classroom pedagogy reforms, alongside curricular changes.

In contrast, the 3v4 comparison yielded significant positive treatment effects after matching, controlling for prior achievement and psychological traits, and including the triple interaction. This suggests that extending the math requirement from three to four years may be more effective in increasing students' confidence in their math abilities. The fourth year of math often consists of more advanced coursework (e.g., Algebra II or Trig+), which may challenge students and reinforce their sense of competence. This aligns with theories of self-efficacy that emphasize the importance of mastery experiences in shaping confidence. Importantly, the treatment effect became more pronounced when gender interactions were included, indicating that average effects were partially masked by gendered trends. Specifically, the significance of Post $\times$ Female in the 3v4 sample suggests that female students experienced unique time dynamics in their math self-efficacy, which confounded simpler models.

Turning to racial heterogeneity, the interaction terms revealed important subgroup differences. In the 2v3 sample, the negative and statistically significant coefficient on Treated × Post × Black suggests that Black students were less likely to benefit from the policy change than their White peers. In the 3v4 sample, I found moderately significant subgroup differences as well, with Hispanic and Asian/PI students showing slightly smaller gains relative to White students. Additionally, when interacting baseline math self-efficacy with race and gender, I found little evidence that students from different backgrounds derived differential benefits from starting high or low in confidence, suggesting that the most important racial and gender differences were structural and not necessarily contingent on prior self-beliefs.

These findings highlight equity concerns. If the benefits of more stringent math requirements are concentrated among already-advantaged students, then such policies risk reinforcing rather than reducing existing gaps. While extending math requirements to four years may boost average self-efficacy, especially for students already on advanced tracks, it is crucial to ensure that all students can access and succeed in these courses. This means pairing curricular mandates with investments in course quality, teacher support, and targeted assistance for students from underrepresented backgrounds.

In summary, math graduation requirement reforms can shape students' beliefs about their abilities, but their effects are uneven. Policymakers should approach such reforms with an equity lens, ensuring that interventions do not simply benefit those already positioned for success. Addressing persistent gender and racial disparities in math confidence may require more than curricular expansion; it may require systemic changes to the classroom experience and school culture.

6.2 Limitations and Future Research Directions

A significant limitation of this study stems from the inability to implement the originally planned staggered difference-in-differences methodology. The ELS:2002 public-use dataset does not include geographic identifiers, which prevented me from aligning students with state-level policy changes. My original approach intended to exploit variation in the timing of math graduation requirement reforms across U.S. states between 2004 and 2009. With access to geographic data, I would have constructed treatment and control groups based on whether and when a student's state increased its math requirements, allowing me to capture policy effects as they unfolded over time in different locations.

This staggered DiD framework would have offered several advantages: more precise estimates of causal effects by leveraging natural experiments, better control for unobserved heterogeneity across states, and the ability to perform placebo tests to assess the parallel trends assumption. Instead, I approximated treatment exposure using students' reports of their school's math requirements and supplemented the analysis with Propensity Score Matching to improve baseline balance. While informative, this design cannot fully account for state-level confounding or policy implementation timing.

A second limitation concerns the structure of the dataset, which only includes math self-efficacy measures at two time points: 10th and 12th grade. This restricts my ability to directly test for parallel trends prior to treatment. Including a lagged self-efficacy measure helps account for baseline differences, but it is not a substitute for observing actual pre-treatment trajectories. This limits causal interpretation, especially in subgroup analyses where time trends may differ.

Third, the outcome of interest, mathematics self-efficacy, is a self-reported psychological belief rather than a long-term behavioral or academic outcome. While I am able to causally link increased self-efficacy to higher STEM career persistence rates in Table 7.3, this is just a variable for intended STEM major evaluated in 12th Grade. As such, future research could extend this and confirm behavioral changes into post-secondary education and beyond.

Finally, the inability to observe local implementation details, such as the content, quality or pace of the additional math coursework, means that I cannot assess whether differences in curriculum or instruction may have mediated the observed effects. It is possible that the nature of the fourth-year math course (e.g., Pre-Calculus versus remedial Algebra) varies significantly across schools and student groups.

Future research should reexamine these questions using datasets with geographic identifiers, enabling precise alignment with state-level reform timelines. This would allow for true staggered DiD estimation and stronger tests of causal assumptions. Studies should also track students beyond high school to evaluate whether gains in self-efficacy translate into measurable shifts in STEM coursework, college pathways, or labor market participation. Finally, greater attention should be paid to differences in implementation fidelity and course quality to assess whether all students, especially those of underrepresented groups, benefit equally from increased graduation requirements.

6.3 Conclusion

This study sheds light on a critical yet often overlooked dimension of education policy: how curricular mandates shape not only what students learn, but what they believe they are capable of. I find that raising high school math graduation requirements from three to four years improves students' self-efficacy in math, a psychological resource long recognized as a key driver of persistence in STEM. But these gains are not evenly shared. The positive effect emerges most clearly once gender is accounted for, suggesting that girls' improvements in math confidence were previously masked by aggregate trends.

This insight is powerful. It means that policy can shift not just performance but perception, particularly for those historically excluded from math-dominated fields. However, it also reveals how deeply gendered these dynamics remain. Girls continue to report lower self-efficacy in math, even when academic performance is held constant. And when I examine interaction effects, I find that gendered patterns of self-belief differ across racial groups, underscoring the importance of an intersectional lens. Addressing gender gaps in STEM cannot mean treating girls as a monolith. Instead, equitable reform must address how gender, race, and academic preparation jointly structure the pathways of students.

As education systems seek to build a more inclusive STEM pipeline, these findings offer both caution and promise. Expanding math access is not a panacea, but it is a start. Policies that assume all students benefit equally risk reinforcing the very disparities they aim to close. To truly empower underrepresented students, especially girls of color, we must look beyond access and toward belonging. That begins with recognizing confidence not as a byproduct of success, but as a condition for it.

CHAPTER 7

APPENDIX

Table 7.1: Ordered Logit Regression Results: 2-Year vs 3-Year Math Requirements

	Coefficient	Std. Error	t value	p value
Treated (3-Year Requirement)	0.512	0.056	9.132	6.73×10^{-20}
Female	0.260	0.052	4.992	5.98×10^{-7}
Black	0.362	0.095	3.816	1.35×10^{-4}
Hispanic	0.206	0.082	2.522	1.17×10^{-2}
Asian / Pacific Islander	0.805	0.101	7.962	1.69×10^{-15}
American Indian / Alaska Native	-0.166	0.272	-0.610	0.542
Multiracial	-0.010	0.123	-0.079	0.937
Parent Education	0.100	0.014	7.111	1.15×10^{-12}
Household Income	0.048	0.013	3.745	1.80×10^{-4}
Urbanicity	-0.342	0.038	-8.917	4.81×10^{-19}
General Challenge	0.426	0.045	9.381	6.57×10^{-21}
Verbal Self-Efficacy	-0.025	0.040	-0.639	0.523
Growth Mindset	-0.002	0.039	-0.051	0.959
Math Score (10th Grade)	0.101	0.004	28.096	1.10×10^{-173}
Verbal Score (10th Grade)	0.028	0.004	6.919	4.56×10^{-12}
Thresholds				
No math course | Pre-algebra/general math	0.873	0.253	3.454	5.53×10^{-4}
Pre-algebra | Algebra I	2.701	0.219	12.355	4.55×10^{-35}
Algebra I | Geometry	3.683	0.217	17.008	7.18×10^{-65}
Geometry | Algebra II	4.900	0.219	22.346	1.32×10^{-110}
Algebra II | Trig / Pre-calc / Calculus	6.937	0.229	30.330	4.63×10^{-202}

Note: The outcome variable is the highest level of math taken in high school, ordered from 1 ("No math course") to 6 ("Trig, Pre-calc, or Calculus"). Thresholds indicate cutpoints between adjacent categories in the ordered logit model.

Table 7.2: Ordered Logit Regression Results: 3-Year vs 4-Year Math Requirements

	Coefficient	Std. Error	t value	p value
Treated (4-Year Requirement)	0.477	0.076	6.319	2.64×10^{-10}
Female	0.225	0.057	3.973	7.10×10^{-5}
Black	0.423	0.094	4.519	6.20×10^{-6}
Hispanic	0.283	0.088	3.222	1.27×10^{-3}
Asian / Pacific Islander	0.781	0.115	6.780	1.20×10^{-11}
American Indian / Alaska Native	0.145	0.297	0.489	0.625
Multiracial	-0.056	0.135	-0.415	0.678
Parent Education	0.113	0.015	7.375	1.65×10^{-13}
Household Income	0.061	0.014	4.361	1.30×10^{-5}
Urbanicity	-0.380	0.041	-9.279	1.71×10^{-20}
General Challenge	0.448	0.049	9.164	4.98×10^{-20}
Verbal Self-Efficacy	-0.041	0.043	-0.947	0.344
Growth Mindset	-0.020	0.042	-0.489	0.625
Math Score (10th Grade)	0.096	0.004	24.770	1.91×10^{-135}
Verbal Score (10th Grade)	0.029	0.004	6.691	2.21×10^{-11}
Thresholds				
No math course | Pre-algebra/general math	0.467	0.277	1.690	0.091
Pre-algebra | Algebra I	2.158	0.236	9.132	6.73×10^{-20}
Algebra I | Geometry	3.067	0.232	13.212	7.46×10^{-40}
Geometry | Algebra II	4.339	0.233	18.584	4.32×10^{-77}
Algebra II | Trig / Pre-calc / Calculus	6.386	0.243	26.291	2.40×10^{-152}

Note: The outcome variable is the highest level of math taken in high school, ordered from 1 ("No math course") to 6 ("Trig, Pre-calc, or Calculus"). Thresholds indicate cutpoints between adjacent categories in the ordered logit model.

Table 7.3: Multinomial Logit Coefficients for Intended Major

Variable	Applied Professions	Environmental Studies	STEM	Undecided	Vocational/Other
Intercept	2.302 (0.289)***	-3.101 (1.205)**	1.165 (0.349)***	1.108 (0.387)**	1.356 (0.350)***
Math Self-Efficacy	0.214 (0.048)***	0.148 (0.183)	0.455 (0.058)***	-0.032 (0.067)	-0.072 (0.059)
Female	0.125 (0.095)	-0.466 (0.356)	-1.149 (0.115)***	-0.358 (0.130)**	-0.329 (0.115)**
Race	0.057 (0.025)*	0.215 (0.122)	-0.028 (0.029)	0.056 (0.035)	0.171 (0.033)***
Parental Education	-0.091 (0.027)***	0.049 (0.107)	-0.033 (0.032)	-0.068 (0.037)	-0.050 (0.033)
Income	0.023 (0.025)	-0.020 (0.100)	-0.023 (0.031)	0.006 (0.035)	-0.003 (0.031)
Math Score (10th Grade)	0.017 (0.006)**	-0.006 (0.025)	0.048 (0.008)***	0.005 (0.009)	-0.013 (0.008)
Verbal Score (10th Grade)	-0.058 (0.008)***	-0.019 (0.030)	-0.060 (0.009)***	-0.046 (0.011)***	-0.039 (0.009)***

Note: Coefficients are relative to the baseline category (Humanities/Social Science). Standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

REFERENCES

Susan Athey and Guido W. Imbens. Design-based analysis in difference-in-differences settings with staggered adoption. *Journal of Econometrics*, 226(1):62–79, 2022. ISSN 0304-4076. doi:https://doi.org/10.1016/j.jeconom.2020.10.012. URL https://www.sciencedirect.com/science/article/pii/S0304407621000488. Annals Issue in Honor of Gary Chamberlain.

Andrew C. Baker, David F. Larcker, and Charles C.Y. Wang. How much should we trust staggered difference-in-differences estimates? *Journal of Financial Economics*, 144(2): 370–395, 2022. ISSN 0304-405X. doi:https://doi.org/10.1016/j.jfineco.2022.01.004. URL https://www.sciencedirect.com/science/article/pii/S0304405X22000204.

Albert Bandura. Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, 84(2):191–215, 1977. doi:10.1037//0033-295X.84.2.191. URL https://doi.org/10.1037//0033-295X.84.2.191.

David Card and A. Abigail Payne. High school choices and the gender gap in stem. Working Paper 23769, National Bureau of Economic Research, September 2017. URL http://www.nber.org/papers/w23769.

Stephen J. Ceci, Donna K. Ginther, Shulamit Kahn, and Wendy M. Williams. Women in academic science: A changing landscape. *Psychological Science in the Public Interest*, 15 (3):75–141, 2014. doi:10.1177/1529100614541236. URL https://doi.org/10.1177/1529100614541236. PMID: 26172066.

Ping Ching Winnie Chan, Tomasz Handler, and Marc Frenette. Gender differences in stem enrolment and graduation: What are the roles of academic performance and preparation? *Economic and Social Reports*, Nov 2021. doi:10.25318/36280001202101100004-eng. URL https://www150.statcan.gc.ca/n1/pub/36-28-0001/2021011/article/00004-eng.htm. Statistics Canada Catalogue no. 36-28-0001.

Joseph R. Cimpian, Sarah T. Lubienski, Jennifer D. Timmer, Martha B. Makowski, and Emily K. Miller. Have gender gaps in math closed? achievement, teacher perceptions, and learning behaviors across two ecls-k cohorts. *AERA Open*, 2(4):2332858416673617, 2016. doi:10.1177/2332858416673617. URL https://doi.org/10.1177/2332858416673617.

Kalena E. Cortes, Joshua S. Goodman, and Takako Nomi. Intensive Math Instruction and Educational Attainment: Long-Run Impacts of Double-Dose Algebra. *Journal of Human Resources*, 50(1):108–158, 2015. URL https://ideas.repec.org/a/uwp/jhriss/v50y2015i1p108-158.html.

Jacquelynne S. Eccles and Ming-Te Wang. What motivates females and males to pursue careers in mathematics and science? *International Journal of Behavioral Development*, 40(2):100–106, 2016. doi:10.1177/0165025415616201. URL https://doi.org/10.1177/0165025415616201.

Lenka Fiala, John Eric Humphries, Juanna Schrøter Joensen, Uditi Karna, John A. List, and Gregory F. Veramendi. How early adolescent skills and preferences shape economics education choices. *AEA Papers and Proceedings*, 112:609–13, May 2022. doi:10.1257/pandp.20221037. URL https://www.aeaweb.org/articles?id=10.1257/pandp.20221037.

Jr. Fryer, Roland G. and Steven D. Levitt. An empirical analysis of the gender gap in mathematics. *American Economic Journal: Applied Economics*, 2(2):210–40, April 2010. doi:10.1257/app.2.2.210. URL https://www.aeaweb.org/articles?id=10.1257/app.2.2.210.

Claudia Goldin. The quiet revolution that transformed women's employment, education, and family. *American Economic Review*, 96(2):1–21, May 2006. doi:10.1257/000282806777212350. URL https://www.aeaweb.org/articles?id=10.1257/000282806777212350.

Joshua Goodman. The labor of division: Returns to compulsory high school math coursework. Working Paper 23063, National Bureau of Economic Research, January 2017. URL http://www.nber.org/papers/w23063.

John Eric Humphries, Juanna Schrøter Joensen, and Gregory F. Veramendi. The gender wage gap: Skills, sorting, and returns. *AEA Papers and Proceedings*, 114:259–64, May 2024. doi:10.1257/pandp.20241026. URL https://www.aeaweb.org/articles?id=10.1257/pandp.20241026.

Alonzo III and Rosa Banda. Cultivating science identity through sources of self-efficacy. *Journal for Multicultural Education*, 10:405–417, 08 2016. doi:10.1108/JME-01-2016-0014.

Ning Jia. Do stricter high school math requirements raise college stem attainment? *Economics of Education Review*, 83:102140, 2021. ISSN 0272-7757. doi:https://doi.org/10.1016/j.econedurev.2021.102140. URL https://www.sciencedirect.com/science/article/pii/S0272775721000595.

Juanna Schrøter Joensen and Helena Skyt Nielsen. Mathematics and gender: Heterogeneity in causes and consequences. *The Economic Journal*, 126(593):1129–1163, 01 2015. ISSN 0013-0133. doi:10.1111/ecoj.12191. URL https://doi.org/10.1111/ecoj.12191.

Kangbok Lee, Yeasung Jeong, Sumin Han, Sunghoon Joo, Junyoung Park, and Kangkang Qi. Difference-in-differences with matching methods in leadership studies: A review and practical guide. *The Leadership Quarterly*, 36(2):101813, 2025. ISSN 1048-9843. doi:https://doi.org/10.1016/j.leaqua.2024.101813. URL https://www.sciencedirect.com/science/article/pii/S1048984324000420.

Allison Mann and Thomas A. DiPrete. Trends in gender segregation in the choice of science and engineering majors. *Social Science Research*, 42(6):1519–1541, 2013. doi:10.1016/j.ssresearch.2013.07.002. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3791309/.

National Science Foundation, National Center for Science and Engineering Statistics. Women, minorities, and persons with disabilities in science and engineering: 2019. `https://www.nsf.gov/statistics/wmpd`, 2019. Special Report NSF 19-304. Alexandria, VA.

Muriel Niederle and Lise Vesterlund. Explaining the gender gap in math test scores: The role of competition. *Journal of Economic Perspectives*, 24(2):129–44, June 2010. doi:10.1257/jep.24.2.129. URL `https://www.aeaweb.org/articles?id=10.1257/jep.24.2.129`.

Lara Perez-Felkner, Kristen Erichsen, Yang Li, Jinjushang Chen, Shouping Hu, Ladanya Ramirez Surmeier, and Chelsea Shore. Computing education interventions to increase gender equity from 2000 to 2020: A systematic literature review. *Review of Educational Research*, 0(0):00346543241241536, 0. doi:10.3102/00346543241241536. URL `https://doi.org/10.3102/00346543241241536`.

Lara Perez-Felkner, Sarah-Kathryn McDonald, Barbara Schneider, and Erin Grogan. Female and male adolescents' subjective orientations to mathematics and the influence of those orientations on postsecondary majors. *Developmental Psychology*, 48(6):1658–1673, November 2012. doi:10.1037/a0027020. URL `https://doi.org/10.1037/a0027020`. Epub 2012 Mar 5.

Lara Perez-Felkner, Samantha Nix, and Krystal Thomas. Gendered pathways: How mathematics ability beliefs shape secondary and postsecondary course and degree field choices. *Frontiers in Psychology*, 8:386, April 2017. doi:10.3389/fpsyg.2017.00386.

Julia B. Smith. Does an extra year make any difference? the impact of early access to algebra on long-term gains in mathematics attainment. *Educational Evaluation and Policy Analysis*, 18(2):141–153, 1996. doi:10.3102/01623737018002141. URL `https://doi.org/10.3102/01623737018002141`.

Elizabeth A. Stuart. Matching methods for causal inference: A review and a look forward. *Statistical Science*, 25(1):1–21, 2010. doi:10.1214/09-STS313.

Jeffrey M Wooldridge. Simple approaches to nonlinear difference-in-differences with panel data. *The Econometrics Journal*, 26(3):C31–C66, 08 2023. ISSN 1368-4221. doi:10.1093/ectj/utad016. URL `https://doi.org/10.1093/ectj/utad016`.

Teng Zhao and Lara Perez-Felkner. Perceived abilities or academic interests? longitudinal high school science and mathematics effects on postsecondary stem outcomes by gender and race. *International Journal of STEM Education*, 9(42), June 2022. doi:10.1186/s40594-022-00356-w. URL `https://doi.org/10.1186/s40594-022-00356-w`.