Essays in the Economics of Education

by

Stephanie Owen

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Public Policy and Economics)
in The University of Michigan
2021

Doctoral Committee:

Professor Susan Dynarski, Chair
Professor Charles Brown
Assistant Professor Sara Heller
Associate Professor Kevin Stange

Stephanie Owen
srowen@umich.edu
ORCID iD 0000-0001-5277-1649

ACKNOWLEDGEMENTS

This dissertation would not exist without the guidance and support of my all-star dissertation committee. Sue Dynarski has modeled how to always foreground real students and policy in the work that we do. Kevin Stange's contagious optimism and detailed feedback have helped keep both me and my research going. I am in awe of Charlie Brown's boundless curiosity and ability to engage with the widest possible range of ideas, not to mention his love of bad jokes. Sara Heller has believed in me and my work at every stage and has gone above and beyond to support me. I have learned so much from her about research, teaching, and mentorship and I intend to pay it forward.

A number of faculty within and outside of the University of Michigan have offered feedback and other forms of support over the past six years. John Bound, while not an official adviser, has attended nearly every presentation I've given and continuously engaged with my research. Jeff Smith, despite leaving Michigan after my second year, has provided informal mentorship and advising from afar. Kathy Michelmore has been a phenomenal supervisor and mentor turned collaborator. I have also benefited from interactions with Martha Bailey, Peter Blair, Sarah Cohodes, Ashley Craig, Mitchell Dudley, Brian Jacob, John Leahy, CJ Libassi, Yesim Orhun, Tanya Rosenblat, Adam Stevenson, Betsey Stevenson, Christina Weiland, Justin Wolfers, and Basit Zafar. I thank Hiba Baghdadi, Kathryn Cardenas, Laura Flak, Julie Heintz, Mim Jones, and Lauren Pulay for their tireless work behind the scenes and patience with my many questions.

My dissertation is the product of several research partnerships. Jasmina Camo-Biogradlija, Julie Monteiro de Castro, Jonathan Hartman, Kyle Kwaiser, Nicole Wagner, and Pam Soltman of the Education Policy Initiative offered guidance with education data and computing resources, administrative support, and financial resources throughout my time at Michigan. I am grateful for the Michigan Department of Education (MDE) and the Center for Educational Performance and Information (CEPI) which provided access to the administrative education records used in Chapter Three. These data were structured and maintained by the Michigan Consortium for Education Research (MCER). MCER data are modified for analysis using rules governed by MCER and are not identical to data collected and maintained by MDE and CEPI. Any opinions, findings, conclusions, or recommendations

expressed in this dissertation are those of the author and do not reflect the view of any other entity. I am grateful to the University of Michigan's Office of Enrollment Management and the Teaching and Learning division of Information and Technology Services for providing data for Chapters One and Two and making it so painless to access. The first chapter of my dissertation (my job market paper) would not have been possible without my fantastic research partners at UM's Center for Academic Innovation, especially Holly Derry, Ben Hayward, Cait Hayward, Tim McKay, and Kyle Schulz. I am so grateful that they took a chance on me and an idea I had for an educational intervention. They have been a delight to work with, and I look forward to continuing our collaboration. Finally, I am appreciative of funding from the Institute of Education Sciences, U.S. Department of Education through PR/Award R305B150012#.

Without a doubt, the best part of my time in graduate school has been the friends I've made and the peers I've had the pleasure of working with. I could not have survived my first year (and second and third...) without Mattan Alalouf and Ellen Stuart. My EPI fam and officemates—Elizabeth Burland, Fernando Furquim, Max Gross, Shawn Martin, Meghan Oster, Shwetha Raghuraman, Michael Ricks, Anna Shapiro, Andrew Simon, and Brittany Vasquez—helped with countless practice presentations, talked through conceptual and technical issues, and helped keep work fun. I never cease to be amazed by the brilliance, kindness, humor, and generosity of classmates like Avery Calkins, Jamie Fogel, Matthew Gross, Sam Haltenhof, Thomas Helgerman, and Dhiren Patki. I'm also in debt to my Ann Arbor friends and roommates for their support and much-needed perspective: Ilana Fischer, Misaki Nozawa, and Jonathan and Joseph Kummerfeld.

Faculty in the economics and math departments at Vassar College introduced me to research and encouraged me to consider graduate school: Kariane Calta, Tracy Jones, Shirley Johnson-Lans, Paul Ruud, and especially Paul Johnson. My supervisors and colleagues at the Brookings Institution and the Urban Institute gave me taste of policy-relevant research and furthered me on my path.

I owe so much of what I have to my wonderful family: Mom, Joe, Dad, Karen, Emily, Ethan, Mark, and Jane. I have never doubted your love and pride in me, and I am incredibly lucky to have a family I both love and like so much. My friends Toby Chaiken, Melissa Cohen, Melissa Ludwig, Alison Russell, and Alana Shein have cheered me on from afar. Finally, Charlotte—the most perfect creature to ever walk the earth—has brought enormous joy during a challenging year-plus of working from home. I can't imagine how I ever lived without her.

I acknowledge my profound privilege and luck in making it to this point, and am aware of the countless talented individuals who have not had the same opportunities. Toni Morrison

used to tell her students, "When you get these jobs that you have been so brilliantly trained for, just remember that your real job is that if you are free, you need to free somebody else. If you have some power, then your job is to empower somebody else." I promise to use my research and privilege to lift others up as I climb.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

ABSTRACT

This dissertation studies the reasons for and consequences of different choices in human capital investment. The theme connecting the three essays is a desire to understand reasons for inequality in educational choices and outcomes, as well as shed light on policies that may or may not remedy those inequalities. In each chapter, I use quantitative casual inference methods and rich administrative data to understand students' educational trajectories.

The first chapter investigates the role of beliefs about academic performance in explaining gender differences in college major choice. I run a randomized controlled trial with undergraduate students across seven STEM disciplines. Treated students receive information about their performance relative to their classmates and to STEM majors. I find that absent intervention, men overestimate their own relative rank by more and are more likely to underestimate how other STEM majors perform, while women are more likely to overestimate others. The intervention shrinks gender gaps in biased beliefs by between a third and half. Treatment also closes the two-credit gender gap in STEM course-taking during the subsequent semester by ten percent. These changes are driven largely by low-performing, overconfident men correctly updating their beliefs and taking fewer STEM credits, rather than encouraging women to stick with STEM.

The second chapter studies the effect of grading policies on college course-taking and major choice, with a focus on differences by gender. I study a natural experiment within the economics department of a large university, which changed its grading policy to give out higher grades in its introductory economics courses. I leverage this variation to compare students with the same underlying performance but who received different letter grades. I find that receiving a higher grade in introductory economics increases the likelihood that a student will take the next course in the sequence by between two and three percentage points, with much smaller effects on economics major choice. Higher economics grades lead more students to declare a major in business—the highest-earning major at the studied institution. I find little evidence that women are more responsive to grades than men. My findings suggest that grade inflation as a policy may work to retain more students within a field, but is unlikely to close gender gaps. Given the complexity of the college major choice and the interdependence across subjects and courses, the optimal policy may not be

straightforward.

In Chapter Three, I study the Advanced Placement (AP) program, which is nearly ubiquitous in American high schools and is often touted as a way to close achievement gaps by income and race. Using administrative data from Michigan, I exploit variation within high schools across time in AP course offerings to identify the causal effect of AP course availability on college choice and degree attainment. I find that higher income students, White and Asian students, and higher-achieving students are both more likely to take advantage of AP courses when they are offered as well as more likely to reap the benefits of taking them. The results imply that not only does the AP program fail to close achievement gaps, it may actually be harming the most disadvantaged students. I find suggestive evidence that the negative effects for low-income and underrepresented minority students are driven by negative spillovers or the diversion of resources from non-AP students and courses, rather than direct effects of these students taking AP.

CHAPTER I

College Field Specialization and Beliefs about Relative Performance: An Experimental Intervention to Understand Gender Gaps in STEM

1.1 Introduction

Understanding how individuals make decisions about college field specialization and how those decisions vary across groups is crucial for educators and other policymakers seeking to address skill shortages in fields such as science, technology, engineering, and mathematics (STEM). National policymakers have called for a dramatic increase in the number of STEM graduates (Olson and Riordan, 2012), and research has documented shortages in certain skills and sectors (Xue and Larson, 2015). In addition to overall shortages, women remain persistently underrepresented in many quantitative fields such as economics, engineering, and computer science. Although they represent more than half of all college graduates, women receive only a third of bachelor's degrees in economics and approximately a fifth of degrees in engineering and computer science (author's calculations using 2017 IPEDS data).

The gender gap in STEM education has implications for both equity and efficiency. The fields with the fewest women also tend to be the highest-paying ones, so differences in field specialization contribute to the gender pay gap. The median lifetime earnings for an economics or computer engineering major—fields where men are overrepresented—are roughly 40 percent higher than that of an English or psychology major—fields where women are overrepresented (Webber, 2019). Furthermore, in a world where individuals specialize according to comparative advantage, removing barriers or frictions that are preventing efficient sorting across fields would increase overall productivity (Hsieh et al., 2019).

While differences in aptitude or performance explain little of the gap in specialization (Cheryan et al., 2017; Ceci et al., 2014), differences by gender in beliefs about performance—conditional on actual performance—may be responsible for differences in educational choices.

Prior empirical work from multiple disciplines has documented systematic differences in men's and women's perceptions of their own performance or competence in various domains and tasks (Niederle and Vesterlund, 2007; Beyer, 1990; Beyer and Bowden, 1997; Lundeberg et al., 1994; Marshman et al., 2018; Vincent-Ruz et al., 2018), while economic theory predicts that beliefs about field-specific ability are a determinant of field specialization (Altonji, 1993; Altonji et al., 2016; Arcidiacono, 2004; Arcidiacono et al., 2016). Research from the lab and the field has shown that information provision can de-bias beliefs and change behavior in a variety of settings (Wozniak et al., 2014; Bobba and Frisancho, 2019; Franco, 2019; Gonzalez, 2017). Several recent field experiments have shown that it is possible to change the academic decisions of college students with light-touch interventions, though cannot disentangle the mechanisms responsible or the reasons for gender differences (Li, 2018; Porter and Serra, 2019; Bayer et al., 2019). Together, these prior strands of work suggest that beliefs about performance may be malleable and salient enough to affect college field specialization choices, but the causal evidence on this mechanism has thus far been limited.

In this paper, I provide the first experimental evidence isolating the effect of beliefs about relative performance on field specialization in college, with an emphasis on understanding differences by gender. I study approximately 5,700 undergraduate students in large introductory STEM courses across seven disciplines at the University of Michigan: biology, chemistry, computer science, economics, engineering, physics, and statistics.[1] The University of Michigan's patterns in STEM degree receipt by gender largely mirror national trends, making it a promising setting to investigate gender gaps. In my primary experimental intervention, I provide students with information about their performance relative to their classmates and relative to STEM majors. In a second treatment arm, I provide a subset of high-performing students with additional encouragement emphasizing their STEM potential.

I collect survey data prior to the intervention and at the end of the semester to measure students' beliefs about relative performance. These data allow me to investigate baseline differences in beliefs by gender independent of any intervention, as well as to understand how the provision of information changes students' beliefs. The size and coverage of my sample allow me to document important heterogeneity in beliefs and belief updating for students at different performance levels, which prior work has largely lacked the power to do. I combine these survey data with administrative data on students' course-taking behavior, my primary short-term measure of field specialization. In the future, I will observe major declaration and degree receipt, as well.

I find that absent any intervention, there are substantial gender differences in two key sets of beliefs about relative performance among control students in the sample. The first is

[1]Throughout Chapter I, references to STEM include economics.

students' prediction of their relative rank in the course. At the beginning of the semester, all students tend to be overconfident in their prediction of their rank, but control men on average overpredict their final performance by 4.5 percentile ranks more than women. Though students become more accurate over the course of the semester, male overconfidence remains. By the end of the term, control men still overestimate their performance by 4 percentiles more than women do; this is due more to overconfidence of low-performing men than underconfidence of women. I also find striking and persistent gender differences in students' accuracy in identifying the median course grade for students who go on to major in STEM. Men are about ten percentage points more likely to think the median course grade for students who go on to major in STEM is lower than it actually is, while women are about 20 percentage points more likely to think it is higher than it is. The patterns in this second type of belief, which no other study has measured, imply male overconfidence and female underconfidence about their performance relative to others. A correlational exercise indicates that these two types of beliefs may account for approximately seven percent of the two-credit (half of a course) gender gap in STEM course-taking in the subsequent semester, even controlling for realized performance and a rich set of academic and demographic characteristics, and explain as much of the gap as prior math achievement.

Providing information on actual relative performance and that of future STEM majors closes the gender gap in beliefs substantially. Among control students, the absolute value of men's error in predicting their own percentile is nearly three percentiles larger than women's; the treatment closes this gap by half. A signed version of this same outcome reveals that overconfident low-performing men correctly update their beliefs downwards, while high-performing men revise upwards. I find no changes in women's beliefs about their class rank, even though they are also inaccurate (though less so than men). The intervention closes the gap in underestimation of the course median for STEM majors by about a third, again by correcting men's beliefs; they are five percentage points less likely to underestimate. The gap in overestimation of the median also closes by nearly a third, this time due to women correctly updating; they are five percentage points less likely to overestimate.

These changes in stated beliefs translate to moderate changes in observed behavior. Providing information closes the two-credit gender gap in STEM course-taking one semester later by ten percent. This appears to be driven exclusively by men, who take three percent fewer STEM credits in the semester following the intervention (though I cannot statistically reject that men and women change their behavior by similar amounts). The results are consistent with low-performing, overconfident men correctly revising their beliefs about their relative performance and taking less STEM as a result; this suggests that absent intervention, men persist in STEM partly because of upwardly biased beliefs about their

relative performance.

Consistent with information provision de-biasing beliefs, students with the most inaccurate pre-intervention beliefs update their beliefs the most and in the direction of the truth. Additional heterogeneity analysis indicates that students we might expect to be on the margin of switching—those already interested in STEM, those who had not yet declared a major, and students earlier in college—changed their beliefs and behavior the most. The intervention does not affect students' effort or performance.

Finally, the results suggest that framing information about relative performance more positively and providing explicit encouragement to continue in STEM is not more effective at changing behavior than information alone for high-performing students. While I find suggestive evidence that high-performing men's beliefs update more positively in response to the information-plus-encouragement intervention compared to pure information, I generally detect no differences by treatment arm, and the effects on course-taking behavior by arm look very similar. For this reason, the majority of the results I present combine the two treatment arms and reflect a general effect of information provision.

This work adds to a number of studies that document systematic differences in men's and women's perceptions of their own performance (Niederle and Vesterlund, 2007; Beyer, 1990; Beyer and Bowden, 1997; Lundeberg et al., 1994; Marshman et al., 2018; Vincent-Ruz et al., 2018; Exley and Kessler, 2019). These studies tend to find that men overestimate their own performance more than women do, at least for domains and tasks considered "male" (Coffman et al., 2019; Bordalo et al., 2019). Prior studies tend to rely on small samples and cannot say much about beliefs over the full distribution of realized performance. I measure relative confidence in two ways: students' beliefs about their own relative rank in their course and their beliefs about how a typical STEM major performs. The first measure confirms previous findings that men are especially overconfident about their performance and sheds light on interesting heterogeneity across the true performance distribution, with the lowest-performing students the most overconfident and the highest performers the most underconfident. No other studies have measured the second type of relative performance belief, which may be especially subject to information frictions and which may be especially salient for specialization decisions. I show that this type of belief in particular has large differences by gender, is strongly correlated with academic behavior, and is moved significantly by the provision of information.

My findings support a long line of economic models theorizing that beliefs about field-specific ability are a determinant of college major choice (Altonji, 1993; Altonji et al., 2016; Arcidiacono, 2004; Arcidiacono et al., 2016), as well as empirical evidence that college students' learning and revision of beliefs about themselves is related to academic decisions

(Stinebrickner and Stinebrickner, 2011, 2012, 2014; Zafar, 2011). These studies combined with documented gender differences in beliefs point to a plausible mechanism for gender differences in STEM (though note that Zafar (2013) finds that differences in beliefs about ability are not a significant determinant of the gender gap in major choice). However, these studies rely on small samples and observational variation in beliefs. Without exogenous variation, the measures of beliefs could be picking up unobserved factors that are the true determinant of behavior. The small sample sizes limit what can be learned about potentially important differences in beliefs for different types of students.

This work also fits in with research from behavioral economics about information provision and belief updating. In lab settings, performance feedback has been shown to close the gender gap in competitiveness (Wozniak et al., 2014; Ertac and Szentes, 2011). In experimental and quasi-experimental work in the field, numerous studies have found that providing individuals with information about their absolute or relative performance changes their subsequent effort and performance (Ashraf et al., 2014; Azmat et al., 2019; Azmat and Iriberri, 2010; Bandiera et al., 2015; Dobrescu et al., 2019; Goulas and Megalokonomou, 2015; Tran and Zeckhauser, 2012). However, the direction of the effect and whether it differs by gender varies across studies. Furthermore, these studies generally do not measure how actual beliefs change, only behavior; combining the two provides more compelling evidence that beliefs influence behavior and by how much. A small handful of studies suggest that providing information about performance can change beliefs and affect outcomes other than performance, such as preferences for academic vs. non-academic high school tracks (Bobba and Frisancho, 2019), college application (Franco, 2019), or enrollment in Advanced Placement courses (Gonzalez, 2017). My study is the first to test this mechanism for college field specialization, and I measure the effects of information provision on beliefs, performance, and subsequent academic choices.

Finally, this work complements several recent interventions that encouraged women to study subjects where they are underrepresented (Li, 2018; Porter and Serra, 2019; Bayer et al., 2019). These studies prove that it is possible to shift course-taking and major choice for women and other underrepresented groups with fairly light-touch interventions. However, due to the designs of these studies, they are not able to isolate mechanisms and compare across groups.[2] Furthermore, they are limited to a single field (economics), meaning we do

[2]The content of Li (2018)'s intervention bundles several mechanisms (information about relative performance, encouragement to major in economics, and information about the field of economics) and varies by student gender and performance. It cannot separately identify the effects of performance information and information about economics for anyone, and cannot separate any of the three mechanisms for high-performing women, who all received encouragement. Porter and Serra (2019)'s intervention involves having recent alumnae visit an undergraduate economics class to talk about their current jobs and the role economics played in their careers. The authors hypothesize that the positive effect on female students is due

not know whether they generalize to other STEM fields and whether they worked by simply shifting students across STEM fields.[3] My study isolates a mechanism—beliefs about relative performance—and compares men and women's beliefs and behavior directly. By including students from seven STEM disciplines, I provide evidence that generalizes beyond economics, and my data allow me to test for substitution across subjects.

Taken together, my experimental results suggest that beliefs about relative performance are a determinant of gender differences in field specialization in college, with male overconfidence the primary force. One-time information provision closed gaps in relative performance beliefs by between a third and a half, and closed gaps in STEM enrollment by ten percent. Though my intervention is low-cost, light-touch, and easily scalable, providing information alone does not eliminate gender gaps. Given how much changes in beliefs seem to correspond to changes in behavior, it would be difficult if not impossible to close the beliefs gap enough to fully close the behavior gap, and further research into other mechanisms is needed. Furthermore, the informational treatment worked by discouraging men rather than encouraging women, which has ambiguous welfare implications for the discouraged men (depending on whether they ultimately change majors and what they choose instead) and their peers (depending on spillover effects of having fewer low-achieving male peers). In future work, I will examine how these short-term effects on course-taking translate to long-term effects on major choice and degree receipt.

The paper proceeds as follows. I introduce my setting and data in Section 1.2, describe my experimental design in Section 1.3, and lay out my empirical methods in Section 1.4. In section 1.5 I document baseline gender differences in beliefs about relative performance. Section 1.6 includes the experimental results of my intervention. Section 1.7 contextualizes the results and Section 1.8 concludes.

1.2 Setting, Data, and Sample

The setting for this study is the University of Michigan - Ann Arbor (UM). UM is considered a highly selective institution (its acceptance rate was 23 percent in 2019) and is the state's flagship. It is a large university, enrolling around 31,000 undergraduate students. I focus on 5,715 undergraduate students enrolled in seven large introductory STEM courses in

to a role model effect, but it could also be due to a previous lack of information about economics-related careers. Since the visiting speakers were all women, they also cannot isolate same-gender effects from general role model effects. Bayer et al. (2019), who sent incoming students welcoming email and information about the field of economics, only target women and underrepresented minorities, so cannot say whether white and Asian men would react similarly.

[3]The exception is Porter and Serra (2019), who test for effects on majoring in other subjects. They find that their intervention pulled women from humanities rather than STEM.

Fall 2019.[4] The courses span seven departments and subjects: biology, chemistry, computer science, economics, engineering, physics, and statistics.[5]

Students in these courses interact with an online platform called ECoach, which is a communication tool designed to provide tailored information and advice to students in large courses. Its intention is to substitute for the personalized one-on-one interactions between instructors and students that are not feasible in courses with hundreds of students. The intervention is delivered through this platform, as are the student surveys.

I use two main sources of data. The first is student administrative records from UM. These data contain all baseline demographic and academic characteristics for the sample such as gender, race, class standing, declared major, standardized test scores, high school GPA, and socioeconomic status. The data also contain students' full academic trajectories while at UM: course-taking, major declaration, official grades, and (eventually) graduation. Because these are administrative data, they contain full information on academic outcomes for all students. Some students are missing information on pre-college characteristics such as high school GPA and parental education, which is collected as part of the application process. This is because some information, such as parental education, is self-reported, and some applicants, such as international and transfer students, do not submit certain information.

The second source is a set of surveys that I administered to all students in the sample at two points in time: one survey before the intervention and one after the intervention. Students took the pre-intervention survey between September and November of 2019, and the post-intervention survey in December.[6] In two of the eight courses (biology and engineering), students received incentives in the form of course credit or extra credit for completing the pre-intervention surveys; an additional four courses (computer science, physics, statistics, and one of the economics sections) received indirect incentives (meaning they needed to complete the pre-intervention survey to access subsequent tasks that offered extra credit). For all courses, taking the pre-intervention survey was a necessary gateway to access most ECoach content.[7] Three courses (biology, computer science, and engineering) offered credit

[4] A second round of the study was planned for the spring semester (referred to as the winter term at the University of Michigan) of 2020. Due to the COVID-19 pandemic and multiple disruptions to the academic and personal lives of students, I canceled the planned second round of the study.

[5] The courses are: Biology 171 (Introductory Biology: Ecology and Evolution), Chemistry 130 (General Chemistry: Macroscopic Investigations and Reaction Principles), Electrical Engineering and Computer Science (EECS) 183 (Elementary Programming Concepts), Economics 101 (Principles of Economics I), Engineering 101 (Introduction to Computers and Programming), Physics 140 (General Physics I), and Statistics 250 (Introduction to Statistics and Data Analysis).

[6] The pre-intervention survey remained open to students throughout the semester, but I drop any responses from after the intervention.

[7] Students who did not respond to the pre-intervention survey could still receive emails sent from ECoach, so not taking the survey did not preclude students from receiving the intervention message.

for the post-intervention survey.

1.3 Experimental Design

1.3.1 Intervention

The intervention consists of two treatment arms, which I refer to as information-only and information-plus-encouragement.[8] The two treatment arms are delivered as online messages and emails to students. The messages were sent a single time in the middle of the semester, at which point students had turned in several assignments and taken at least one exam. The messages were timed to align with the beginning of course selection and registration for the subsequent semester.

In the first treatment arm, which I refer to as the information-only intervention, I provide students with information about their performance relative to their classmates and relative to STEM majors. Specifically, the message includes a histogram showing the current distribution of grades in the course. Their own grade is highlighted and their percentile is labeled (e.g., "You're at the 75th percentile"). The graph also includes a call-out informing students about the typical grade in the course for a STEM major (e.g., "STEM major median: B+"). All of the key information in the chart—the student's score and percentile and the median for STEM majors— is repeated later in the message. The second part of the message gives further context about grades in the course, listing the course median for all students, students who go on to major in the field associated with the course, and (again) students who go on to major in STEM.[9] The final part of the message includes a list of links to set up advising appointments in various STEM departments (with the department the course is in appearing first). Appendix Figure A.1 shows an example of an information-only message.

The second treatment arm, which I refer to as information-plus-encouragement, was sent to a random subset of high-performing students, defined as those performing above the course median at the time of randomization. It includes all of the same information as the information-only intervention. However, it is framed in more positive language calling attention to the student's strong performance ("You're performing like a STEM major!"

[8]This study was pre-registered with the American Economic Association's registry for randomized controlled trials under RCT ID AEARCTR-0004644: https://doi.org/10.1257/rct.4644-1.0.

[9]For biology, economics, computer science, and engineering, the associated major is just the field. For classes where fewer than 10 percent of students go on to major in the subject, the message emphasizes multiple majors. The physics and chemistry courses tend to serve many more engineering majors than physics or chemistry, so the associated major is the subject *or* engineering. The statistics course serves students who ultimately major in many fields, so the associated major is statistics, economics, or computer science—the most common STEM majors for students who take the course.

rather than "Here's how you're doing") and includes language explicitly encouraging the student to consider or stay in a STEM major. (Based on the student's response to the pre-intervention survey item about their intended major, they are either urged to stay in their major or to consider a STEM major.) Appendix Figure A.2 shows an example of an information-plus-encouragement message.

In designing a second treatment arm, I wanted to test whether the framing of the information affected how students incorporated it. The findings of Li (2018), an experimental intervention that bundled relative performance information with encouragement and information about the field of economics, suggest that the encouragement aspect may be important for high-performing women in particular but cannot disentangle the various components.[10]

Notably, students already know (or can easily see in multiple sources) their score in the course, but generally are not told their exact percentile. Information about historical course medians is available through an online system maintained by the university, but this system reports only overall course medians and not medians specific to certain populations like STEM majors. Furthermore, evidence from the pre-intervention survey suggests that students do not have accurate beliefs even about the information that is readily available; less than a third of students accurately identified the historical course median.

Students in the control condition also received messages reminding them of their current score, but containing no additional information about their relative performance. The control messages reminded students that course registration for the next semester was soon and contained the same advising links. I sent control messages to limit any confusion or spillover among control students; the intention was that they would not wonder why they did not also receive a message about their grades. Appendix Figure A.3 shows an example of a control message.

1.3.2 Treatment Assignment

I assign treatment status at the student level, stratified by course, gender, and performance at the time of randomization (above versus below the course median). This results in $8 \times 2 \times 2 = 32$ strata.[11] Within each of the 16 below-median strata, the probability of receiving the information-only treatment is 0.5. Students who are above

[10] Li (2018)'s intervention had a positive effect on high-performing women, who received relative performance information, encouragement to major in economics, and information about the field of economics; it cannot identify which of the three mechanisms worked. Men did not receive any encouragement, so the study also cannot say whether men and women respond differently to encouragement.

[11] Though there are seven courses with multiple sections each, the two economics sections operate independently (notably for grading), so I consider them separately for randomization.

the median are eligible for the second treatment arm; within the 16 above-median strata, the information-only and information-plus-encouragement treatment are each assigned with probability 1/3. I chose these treatment probabilities to maximize statistical power across the main and subgroup comparisons I was most interested in. To achieve a balanced sample in practice and not just in expectation, I re-randomize until each pre-treatment characteristic is balanced within strata (minimum p-value of 0.1). I account for this re-randomization and its implications for inference in my analysis by using randomization-based inference. This method resulted in 2,382 control students, 2,393 students who received the information-only treatment, and 940 who received information plus encouragement. Figure 1.1 summarizes the experimental design.

Fifteen percent of the sample are enrolled in more than one of the included STEM courses. To account for this, I randomly choose (with equal probability) which of their courses they will be considered in for the experiment. Within that course, they are assigned to a treatment condition like everyone else. For their other courses, they receive no message (not even a control message).

1.3.3 Sample Characteristics and Balance

Table 1.1 shows demographic and academic characteristics for the sample by treatment status, based on university administrative data. This table also tests for balance on pre-treatment characteristics between control students and treated students. (Table 1.1 pools students receiving either treatment; a balance table that separates the two treatment arms is presented in Appendix Table A.1. I also test for balance separately by gender in Appendix Table A.2).

The total experimental sample includes 5,715 students, of whom slightly under half (48 percent) are women. The majority of students (55 percent) are white. A large proportion (27 percent) are Asian, while smaller numbers identify as non-Black Hispanic (seven percent) or Black (three percent). This largely reflects the demographics of the university, though white and particularly Asian students are even more overrepresented in these STEM courses compared to the university as a whole. The majority of students are in their first or second year of college (42 and 40 percent, respectively).[12] The average UM student and the average student in this sample come from a socioeconomically advantaged background: 60.5 percent have a parent with a graduate or professional degree, and only 15 percent are first-generation (meaning neither parent has a bachelor's degree). The majority (64 percent) have family

[12]Technically, UM measures class standing based on credits accumulated, so that, for example, some students classified as sophomores may be first years with enough credit (from previous courses, transfer, AP, etc.) to count as sophomores.

incomes above $100,000. Roughly half of the sample (52 percent) are Michigan residents.

The average cumulative GPA while at UM is 3.41 (students in their first semester do not yet have values for this variable). UM is a highly selective school, and this is reflected in the high average test scores (e.g., 710 out of 800 on the SAT quantitative section) and high school GPA (3.88 average). A large majority (83 percent) took calculus in high school. At the time of randomization, the majority of students (56 percent) had not officially declared a major. Of those who had declared, most were engineering majors (23 percent of the full sample). Nine percent were in a non-engineering STEM major, and 11 percent had declared a non-STEM major.[13]

I test for balance on each pre-treatment characteristic, as well as for the proportion of students missing information on each characteristic, with a regression of the characteristic on treatment status, controlling for strata. I find one significant difference out of 36 tests, fewer than would be expected by chance. Treated students have an average ACT reading subscore that is 0.1 points lower on the 36-point scale, which is substantively small. I also test for whether the characteristics jointly predict treatment status, again controlling for strata; the p-value from this F-test is 0.840.

Though not shown in Table 1.1, the highest proportion of students are in the statistics and chemistry courses (26.9 and 19.7 percent, respectively), and the lowest number are in engineering and physics (7.9 and 5.7 percent, respectively); these proportions reflect differing enrollments and course sizes. The full breakdown of the sample by course and gender is available as Appendix Table A.3. Fifteen percent of students are enrolled in more than one of the seven courses, but are only considered in the experimental sample for a single course, which is chosen randomly (see section 1.3.2).

1.3.4 Take-up

Students could receive the intervention in two ways. The first was an email that was sent directly to their official university account. The second was from within ECoach, which students can visit at any time to view relevant information and other messages about the course. There were some minor formatting differences, but the content of these two formats—including the visual element, the histogram—was identical.

Among students who were sent a treatment message, 83 percent viewed it in some format. 57.5 percent viewed the message only as an email, three percent saw the message only within ECoach, and 23 percent viewed it in both formats. Women were more likely to view the

[13]Engineering is its own college and prospective engineers are admitted directly into the program as incoming first years. Many eventual science, humanities, social science, and other popular majors appear as undeclared during their first and second year, until they meet major prerequisites and apply for the major.

message (in either form) than men: 85.5 percent of women compared to 81.2 percent of men ($p = 0.001$).

I further examine whether certain types of students were more likely to read the intervention messages by regressing receipt of the message (in any form) on all pre-treatment characteristics, as well as the course the student is in and whether they were performing above the course median (included as Appendix Table A.4). Conditional on all other covariates, women, high-performing students, Black students, and those in the statistics, computer science, biology, and engineering courses were most likely to view the messages.

1.3.5 Survey Response

Around three quarters of students responded to the pre-intervention survey, while slightly less than half (48.7 percent) responded to the post-intervention survey. Women were seven percentage points more likely to respond to each survey than men ($p < 0.001$). I test for differential survey response by treatment status and find none. I show item-level response rates for the items used in my analysis as Appendix Table A.5. The item-level response rates to the post-intervention survey range from 41.3 percent (for beliefs about own performance) to 46.6 (for intended major).

I more thoroughly test for differences in survey response by pre-treatment characteristics in Appendix Table A.6.[14] I do this by regressing an indicator for post-intervention survey response on the full set of observed pre-treatment characteristics. Similar to the unconditional difference, women were seven percentage points more likely to respond to the post-treatment survey. Higher-performing students (those in the top half of their course at the time of randomization) also had a seven percentage point higher response rate; the gender-by- performance interaction is not significant. Students with higher college and high school GPA also responded at higher rates. Students in the statistics and engineering courses have the highest response rates; recall that instructors in these courses offered extra credit for both surveys. Similarly, students declared as engineering majors were more likely to respond than any other major. The courses with the lowest response (the first economics section, which is the omitted category, and chemistry) had generally low engagement with ECoach. Younger students (first years and sophomores) were more likely to respond than upper-year students. This is consistent with students who are missing a measure of prior college GPA being more likely to respond, as this generally indicates that they are in their first semester of college. Asian students had the highest response rates: seven percentage points more than the reference group of American Indian or multiple race students. Finally,

[14] I focus on the post-intervention survey here, since I estimate treatment effects on post-intervention variables.

students missing an SAT score are 30 points less likely to respond, while students missing a value for high school GPA are 48 points more likely. These are somewhat hard to interpret because students can be missing values for multiple reasons. Missing scores may indicate international or transfer students; additionally, the state of Michigan switched from requiring 11th graders to take the SAT rather than the ACT in 2016, so having one score over another may indicate cohort.

It is not surprising that different types of students were more or less likely to respond to the surveys. Survey response is independent of estimated treatment effects on my primary outcomes, which use administrative data, but could affect the internal and external validity of analyses using survey outcomes. To asses internal validity of analysis using survey outcomes, I run the same balance tests as in Section 1.3.3, this time conditional on responding to the post-intervention survey. These results, shown in Appendix Table A.7, indicate that all pre-treatment characteristics remain balanced when I limit to survey respondents (p-value from joint F-test = 0.953). The other potential concern is that any analysis done using survey data does not generalize to the full sample. To address this, I run two robustness checks. In the first, I estimate treatment effects on administrative data outcomes using only the sample who responded to the survey. In the second, I re-estimate effects on survey outcomes using inverse probability weighting to make survey respondents resemble the full sample on their observable characteristics. In both cases, the point estimates are similar.

1.4 Empirical Method

1.4.1 Method for Descriptive Analysis

As motivating evidence for the hypothesis that gender differences in beliefs in relative performance are responsible for some of the gender gap in field specialization in college, my first set of results are a descriptive analysis of students' beliefs and how those beliefs are related to behavior. For the descriptive analysis, I restrict the sample to students assigned to the control condition to measure beliefs in the absence of any intervention. I use responses to the pre- and post-intervention surveys to understand how students update their beliefs over the course of a semester. I use the administrative transcript data to study how beliefs correlate with observed behavior. In all descriptive analyses, I limit the sample to control students who responded to both surveys to avoid any confounding changes due to differential response over time.

1.4.2 Method for Estimating Treatment Effects

To estimate the main effect of the intervention, I use the full sample of students and estimate the following specification:

$$Y_i = \beta_0 + \beta_1 Treat_i + \boldsymbol{\gamma X_i'} + \delta_s + \epsilon_i \tag{1.1}$$

where $Treat_i$ indicates assignment to the either treatment, X_i is a vector of pre-treatment covariates (everything listed in Table 1.1), and δ_s are dummy variables for all but one of of the 32 gender-by-course-by-above-median strata.[15] In this specification, β_1 is the estimated intent-to-treat (ITT) effect, or the effect of being sent an intervention message, for all students. Scaling the ITT by the inverse of the message take-up rate ($1/0.83 = 1.2$) gives the effect of treatment on treated students (TOT).

I am particularly interested in how the treatment differentially affects men versus women. To estimate effects by gender, I add in an interaction for female students:

$$Y_i = \beta_0 + \beta_1 Female_i + \beta_2 Treat_i + \beta_3 Female_i \cdot Treat_i + \boldsymbol{\gamma X_i'} + \delta_s + \epsilon_i \tag{1.2}$$

Here, β_2 gives the treatment effect for men, and $\beta_2 + \beta_3$ gives the effect for women.

In most reported results, I pool the two treatment arms together and estimate a single treatment effect. The estimated treatment effects are therefore an average of the information-only and information-plus-encouragement treatments. To separately estimate and compare effects of the two treatment arms, I limit the sample to above-median students, who were eligible for the second treatment arm, and estimate:

$$Y_i = \beta_0 + \beta_1 Info_i + \beta_2 Encourage_i + \boldsymbol{\gamma X_i'} + \delta_s + \epsilon_i \tag{1.3}$$

where $Info_i$ indicates assignment to the information-only treatment, $Encourage_i$ indicates assignment to the information-plus-encouragement treatment, and everything else is as above. I also estimate the effect of the two treatment arms by gender and with a specification analogous to Equation 1.2 (where I include indicators for each treatment and interactions between each treatment and gender).

In all analyses, I estimate ITT effects, or the effect of being sent an intervention message. I estimate treatment effects on students' beliefs about their relative performance using outcomes measured in the post-intervention survey. I estimate effects on short-term measures of field specialization (course-taking in the semester following the intervention) based on

[15] I also report estimates without covariates in the appendix.

administrative transcript data. I investigate additional mechanisms using outcomes and characteristics collected in the survey and available in administrative data.

All tables report robust standard errors and significance levels. In addition to standard inference, I also calculate p-values using randomization-based inference. In this approach, randomness in estimates comes from assignment of a fixed number of units (students) to treatment, rather than from sampling from a population. To implement, I re-assign treatment status 10,000 times, using the same procedure used in the original randomization. This accounts for the fact that my re-randomization procedure changes the distribution of test statistics, because I discard any re-randomizations that do not meet the pre-specified balance rule (Bruhn and McKenzie, 2009). Randomization inference also addresses concerns about clustered data, because it preserves the underlying data structure, including any mean or higher-order correlations. Under each "treatment" assignment, I calculate a test statistic of interest (a main effect, the effect for men, the effect for women, or the differential effect). This process generates a distribution of potential treatment effects that could be due to baseline differences between students assigned to treatment and control. (Note that this accounts for any outliers that may be driving treatment effects.) For each effect, I calculate the share of the 10,000 simulated estimates that are larger in absolute value than the estimate observed under the true treatment assignment; this proportion represents the randomization-based p-value. Note that while the traditional sampling approach tests a null hypothesis of no average effect, randomization inference tests a sharp null hypothesis of no effect for any individual. A comparison of sampling or model-based p-values and randomization-based p-values is presented in Appendix Table A.8. Although they represent different conceptual approaches, the model- and randomization- based p-values produce virtually identical conclusions.

I address concerns of data mining and the possibility of finding falsely significant results in two ways. First and most importantly, this project was pre-registered with the American Economic Association's trial registry, and I pre-specified all experimental analyses. Any exploratory, non-pre-specified analyses are identified as such. To further test the robustness of the main results to testing multiple hypotheses, I implement two types of adjustments. The first approach controls the false discovery rate (FDR), or the proportion of null-hypothesis rejections that are Type I errors. I implement the simple procedure in Benjamini and Hochberg (1995) as well as the two-stage procedure from Benjamini et al. (2006). The second approach controls the family-wise error rate (FWER), or the probability that at least one of the true null hypotheses in a family of hypothesis tests is rejected, using the permutation resampling method in Westfall and Young (1993). Appendix Table A.9 compares unadjusted p-values to the inferences from these three methods. The inferences about the statistical

significance of the main results generally hold up under these adjustments, with the findings on men's beliefs and behavior in particular surviving at conventional significance levels.

1.5 Control Students' Beliefs about Relative Performance

I begin by documenting systematic gender differences in students' beliefs about their relative performance, using responses to the student surveys at two points in time. This section focuses on control students only, to understand beliefs in the absence of intervention. This analysis is further limited to students who respond to both surveys, to avoid any confounding changes due to differential response over time. After showing raw differences in the beliefs of women and men, I perform a decomposition exercise to see how much these differences correspond to differences in course-taking.

1.5.1 Student Beliefs about Their Own Percentile

I measure baseline beliefs about relative performance in two ways. The first is how accurately students perceive their own relative rank in the course, measured by comparing what they predict their final percentile will be (in the pre- and post-intervention surveys) to their true percentile at the end of the course.[16] I do this at two points in time to see how beliefs change over the course of semester. I show this visually and also report the average errors in beliefs; I report both the absolute value of the error as well as a signed error to convey whether certain groups are over- or underestimating their performance.

Control students begin the semester inaccurately predicting their performance.[17] The average control student overpredicts their performance by 15.9 percentile ranks, meaning they expect to perform considerably better than they actually do. Because some students underpredict their performance (a negative error), the average absolute value of a student's error is even larger in magnitude: 28 percentile ranks. There are significant differences by gender and performance. The average man assigned to the control condition overpredicts his final performance by 18.3 percentiles, while the average woman overpredicts by 13.5 ($p<0.05$). Low-performing (below-median) students tend to overestimate their performance (by 30.3

[16]The survey item asks students to fill in a value from 1 to 100: "In terms of my final grade, I expect I will do better than ____% of my classmates in [*course*]." This survey item is not incentive-compatible, meaning students are not incentivized to give an accurate prediction. Note that doing so would itself constitute a treatment and could cause students to update their beliefs. The fact that control students nonetheless update reported beliefs over time suggests that the responses capture real beliefs despite not being incentivized.

[17]Recall that students responded to the pre-intervention survey between September and November. Over 80 percent responded in September and nearly 90 percent took the first survey before the first exam in their course. When first asked to predict their performance, they would have had limited performance feedback from assignments.

percentiles), while high-performing ones tend to underestimate, though to a lesser extent (average underprediction of 2.7 points). Low-performing men are the most overconfident (overpredicting by an average of 34.5 percentiles, compared to 27.7 for low-performing women) while high-performing women are the most underconfident (underpredicting by an average of 5.8 percentiles compared to less than a percentile for high-performing men).[18]

Panel (a) of Figure 1.2 summarizes how accurate control students' beginning-of-semester predictions of their relative performance are by gender and realized performance. This graph plots realized performance (percentile rank in terms of final course grade) against predicted performance, grouping students into 50 equally sized bins by gender (roughly ten students per bin); the x- and y-values are the within-bin means. The fact that most plotted points fall above the 45-degree line confirms visually that most students start the semester overpredicting how they will do. The graph also makes clear that the lowest-performing students are the most overconfident, while the highest performers are the most underconfident. What is striking is that men's beliefs are consistently higher than the beliefs of women performing equally well. I formally test this in a regression of predicted percentile on true percentile, gender, and their interaction. The intercept for women is approximately eight percentiles lower, while the slopes are indistinguishable. The flatness of the slopes is consistent with students largely guessing (or not caring about) how they will do, but the gender differences suggest some underlying difference in the process of predicting.

Even absent intervention, we would expect students to update their beliefs over the course of the semester as they learn about their performance through exams, assignments, and other feedback. At the end of the semester (right before final exams), control students' predictions are more accurate than they were at the beginning. The average student is still overpredicting, but by less: 4.7 percentiles compared to 15.9 at the start of the semester. Compared to an absolute value error of 28 percentiles at the beginning of the semester, the average control student's absolute error at the end of the semester is 19.2. The fact that the change in the signed error is similar to the change in the absolute value of the error suggests that it is primarily the students who were initially overpredicting who updated. Though both men and women have updated, a gender gap in beliefs remains: the average man assigned to the control condition overpredicts his final performance by 6.7 percentiles, while the average woman overpredicts by 2.7. The gender gap among low-performing students is only slightly smaller compared to the beginning of the semester: below-median men are 5.7 percentiles more overconfident than women (15 vs. 9.3). The gender gap among high-performing students has shrunk to 4.1 percentile points ($p<0.1$).

[18] Whenever I group students by high-performing (above-median) and low-performing (below-median), I use performance measured in the middle of the semester, at the time of randomization.

These changes are reflected in Panel (b) of Figure 1.2. The plotted points are now clustered closer to the 45-degree line, and the points on the left (i.e., the lower performing students) shift more over the semester; this means that students became more accurate, particularly the ones who were previously the most overconfident. While the beliefs gap between the highest performing men and women has closed over time, lower performing control men remain more overconfident than women performing similarly.

1.5.2 Student Beliefs about Other STEM Majors

My second measure of beliefs about relative performance focuses on what students believe about STEM majors. I ask students what they think the median grade in their course is among students who go on to major in a STEM field; I can then compare their answers to the true median.[19] This measure captures how difficult students perceive the course to be, how well they think they must do to pursue STEM, and (implicitly) how they compare to other STEM majors.

Panel (a) of Figure 1.3 summarizes how well students can identify the STEM major course median at the beginning of the semester, by gender. (I again limit the sample to control students who also answered the analogous end-of-semester survey item.) At the outset of the course, 33 percent of men and 27 percent of women accurately report the median. Men are much more likely to underestimate the median (30 vs 19 percent), while women are much more likely to overestimate (53 vs 36 percent). Note that in this case, underestimating means a student thinks their (potential) peers are doing worse than they actually are. In other words, this suggests that women may believe the bar for majoring in STEM to be higher than men do.

Control students' beliefs about the course median for STEM majors change little over the semester (Panel (b) of Figure 1.3). This is unsurprising; though they learn about their own performance and, to a lesser extent, that of their peers, they receive no direct information about STEM majors' grades in particular. By December, when they respond to the post-intervention survey, 26 percent of control men and 17 percent of control women underestimate the median; 36 percent of men and 55 percent of women overestimate. Low-performing men are the most likely to underestimate the median (32 percent), while high-performing women are the most likely to overestimate (69 percent).

[19]The survey item asked, "When thinking just about students who declare a major in math, science, engineering, or economics, what do you think was their median grade in [*course*]?" The true course medians for STEM majors for the seven courses are: B for Biology, Chemistry, and Physics; B+ for Economics and Statistics; and A- for Engineering and EECS. I calculate these using historical administrative data on students who took each course in the 2014-15, 2015-16, or 2016-17 academic year and who declared a STEM major within three terms of taking the course.

Students also responded to questions about their beliefs on the overall course median and the course median for students who major in the subject affiliated with the course (e.g., the Econ 101 median among students who declare an economics major). Beliefs about the median grade for subject majors are similar to beliefs about STEM majors. For beliefs about the overall course median, all students are much more likely to underestimate, but the differences by gender are much smaller. Among control men, 55 percent underestimate, 33 percent are accurate, and 12 percent overestimate the overall median at the end of the semester. Among control women, the proportions are 50, 35, and 15 percent. The negligible gender differences in overall median beliefs imply that it is not the case that men and women have different beliefs about grades or grade inflation generally. Rather, they hold different beliefs about the selection into STEM, with women setting the bar for STEM higher.

1.5.3 Beliefs about Relative Performance and Course-taking: A Correlational Exercise Using Control Students

In the previous section, I find that men are more overconfident than women about their own place in the course distribution, even by the end of the semester when they have nearly full information about their performance; this is especially true for lower performing men. Men are also more likely to underestimate how STEM majors perform, while women are much more likely to overestimate. These two sets of findings about students' beliefs—about their own relative rank and about the performance of other STEM majors—work in the same direction, and suggest a story of relative male overconfidence and female underconfidence.

This may be part of the explanation for differential rates of STEM enrollment and persistence. In the semester following the course, control men took an average of two STEM credits more than women ($p<0.001$). (A single STEM course is usually four credits, so this represents half of a course.) This is consistent with men being more confident than women about their performance and confidence affecting course-taking. While suggestive, this relationship is correlational and does not account for the myriad factors which may differ by gender.

To investigate more systematically whether beliefs about relative performance are related to the gender gap in course-taking, I perform a decomposition following Gelbach (2016). This accounting exercise uses the omitted variable bias formula to partial out how much the addition of a variable to a regression changes some base coefficient—in my case, the coefficient on female, which represents the gender gap. An advantage of this approach relative to one that progressively adds covariates is that it is not sensitive to the order in which covariates are added.[20]

[20]The Gelbach decomposition is conceptually similar to a Kitagawa-Oaxaca-Blinder decomposition, and

I apply the decomposition to a model where I regress the number of STEM credits in the semester following the course on a female dummy, all of the demographic and academic controls in Table 1.1, the student's final percentile rank in the course, their prediction of their final percentile, and dummies for whether they are under- or overestimating the median course grade for STEM majors. The results are presented in Table 1.2. Only control students who responded to both surveys are included in this exercise.

The full set of belief, performance, academic, and demographic variables account for roughly half of the observed gender gap in credits (2.15 credits in this sample). A student's declared major when they took the course explains by far the largest part of the gap: 32 percent. A student's score on the math placement test they take upon entering UM explains an additional seven percent of the total gap. Demographics, high school and college achievement, and student level together explain three percent. Students' beliefs about their own course percentile explain around two percent of the gender gap in credits, and beliefs about the course median for STEM majors explain an additional 5 percent. Together, the beliefs measures account for seven percent of the total gender gap, and 14 percent of the part of the gender gap that is explained by covariates. The decomposition suggests that students' beliefs about other STEM majors may be particularly important.

My results thus far demonstrate that women and men have systematically different beliefs about their relative performance in STEM courses, and that even conditional on true performance and a rich set of academic and demographic covariates, these beliefs are related to the gap in field specialization in college. My study is one of very few that can connect beliefs about consequential real-world performance to observed, real-world outcomes, and the largest-scale study in the context of postsecondary specialization. Furthermore, I show that students' beliefs about the performance of other STEM majors is consequential for the STEM behavior gap; no other studies have measured this belief, which may be particularly subject to information frictions and particularly salient for specialization decisions.

Even accounting for a rich set of controls, this relationship remains correlational. It is possible that my measures of beliefs may be picking up some omitted factor that is actually responsible for behavior, and correlations between the covariates make the magnitudes hard to interpret. To isolate the causal role of relative performance beliefs, my experiment attempts to exogenously change beliefs and study how academic decisions change as a result.

in fact is equivalent once interactions between the covariates and gender are included.

1.6 Experimental Results

1.6.1 Effect of Intervention on Student Beliefs

I begin by estimating treatment effects on students' beliefs, using measures of relative performance beliefs similar to those described in Section 1.5. The first measures the accuracy of students' beliefs about their own relative performance by subtracting the student's true percentile from what they estimate their percentile to be at the end of the semester. Here, I use mid-semester performance as the realized percentile, because end-of-semester performance could itself be affected by the intervention if students adjust their effort. (For this reason, the control means in the treatment effects tables differ from the values reported in Section 1.5.1.) I test for effects on performance directly in Section 1.6.3.[21] I report both an absolute value measure as well as a signed measure that captures the direction of the error. Second, I measure the accuracy of beliefs about the performance of STEM majors by creating two indicator variables for whether a student is over- or underestimating the course median for students who go on to major in STEM.

Table 1.3 shows treatment effects on beliefs outcomes, for the full sample as well as separately for men and women.[22] As I show later, I do not find strong evidence of differential effects on beliefs or behavior by treatment arm, so in this table I combine the two treatment arms. All treated students received the same informational content; the only difference between the arms was whether the information was framed in a neutral or positive way.

The results for the absolute value of the error in the student's predicted percentile indicate that the average student correctly updates their prediction by approximately 1.5 percentiles. (A negative treatment effect means the error is getting smaller.) This appears to be driven by men updating: they correct their beliefs by 2.2 percentiles, while women's absolute error shrinks by a statistically insignificant 0.7 percentiles (though note I cannot reject that men and women's beliefs change by the same magnitude). The gender gap in this measure among control students is 2.7 percentiles (20.3 for men minus 17.6 for women), so the covariate-adjusted gap in the absolute value prediction closes by half.

When I look instead at the signed error in percentile beliefs, I find no average treatment effect overall or for either gender. However, the fact that the absolute value of the error changes implies that this null finding is masking belief updating that goes in both directions.

[21] I also estimate treatment effects on a version of the percentile belief outcome where I use final performance rather than mid-semester performance as the realized performance (not shown). The signs are similar but the magnitudes somewhat smaller. This is not surprising given that the intervention told students their mid-semester percentile; they updated their beliefs in the direction of the signal they received.

[22] Treatment effects on beliefs outcomes estimated without covariates are included as Appendix Table A.10. The point estimates are very similar.

I explore this further below.

The estimated effects on students' beliefs about the median course grade for STEM majors indicate that the intervention also closed part of the gender gap in this second type of belief. Receiving the informational intervention made men 5.2 percentage points less likely to underestimate the median and made women 5.1 percentage points less likely to overestimate. The gender gap in underestimating among control students is 9.8 percentage points (with men more likely to underestimate) and the control gap in overestimating is 17.7 percentage points (with women more likely to overestimate). Comparing control and treatment gender gaps, the treatment closes the gap in both measures by roughly a third. Both changes suggest that men are becoming less overconfident relative to women.

In Table 1.4 I further disaggregate students by whether they were below or above the course median at the time of treatment. Recall that lower-performing (below-median) control men were particularly overconfident in both types of beliefs and higher-performing (above-median) control women were particularly underconfident in terms of the STEM median measure. If the groups who were the most inaccurate correctly revise, we would expect the point estimates on percentile beliefs and underestimating the median to be negative for low-performing men, and the point estimates on overestimating the median to be negative for high-performing women. Table 1.4 also separately estimates effects of the two treatment arms for the above-median students; only above-median students were eligible for the second treatment arm of information paired with encouragement.

I find that students' beliefs about their own percentile change in the expected direction. Low-performing men, who in the absence of intervention overestimate their percentile by 21.4 percentiles, update downwards by 3.7 percentiles. High-performing men show the opposite pattern: they underestimate their percentile by seven points absent the intervention, but receiving either treatment (pooled effect) causes them to update upwards by four percentiles. In other words, both low and high-performing men become more accurate in their predictions.

I find that low-performing men, who are most likely to underestimate the course median for STEM majors without the intervention, become 8.8 percentage points less likely to do so (a change of 28 percent relative to the control mean of 31.8); I detect no change for any other group. Similarly, high-performing women, who are most likely to overestimate the median, see the largest change in that measure. The pooled estimate suggests the intervention makes high-performing women 11.5 percentage points less likely to overestimate (a change of 17 percent relative to the control mean of 68.1).

I find limited evidence that the encouragement treatment arm was more effective than the purely informational treatment for high-performing students. The point estimates for above-median men suggest that the information-plus-encouragement message may

have led to a larger positive update in percentile beliefs for this group, but it is only marginally significantly different from the information-only effect (p = 0.085). The effects of encouragement for high-performing women are also larger than those of pure information (1.3 vs. 0.4 percentiles) but not statistically different. Overall, my results do not provide strong support for a differential treatment effect, so for the remainder of the paper I combine the treatment arms and consider the effect of receiving any type of informational treatment. (I discuss estimated effects by treatment arm on my primary outcomes in the next section and show them in Appendix Table A.12.)

1.6.2 Effect of Intervention on STEM Persistence

My primary behavioral outcome is STEM persistence, which I operationalize as the number of credits a student attempted in the semester following the intervention, as well as a binary indicator for taking any STEM courses. I classify courses by two-digit Classification of Educational Program (CIP) code, developed and maintained by the U.S. Department of Education's National Center for Education Statistics.[23] The following subjects (CIP codes) are considered STEM: natural resources and conservation (03), computer and information sciences (11), engineering (14), biological and biomedical sciences (26), mathematics and statistics (27), physical sciences (40), and economics (45.06; see footnote). This outcome comes from the administrative data; attrition or missingness occurs only if a student graduates or drops out. If a student does not show up in the data in a given term, I code them as taking zero credits and courses.[24] In future work, once more time has passed, I will examine major declaration (medium-term) and STEM degree attainment (long-term). I use additional survey outcomes and a prediction exercise to estimate how the observed short-term effects are likely to translate into long-term effects.

Table 1.5 reports estimated treatment effects on STEM persistence in the semester following the intervention.[25] The first column shows that the average effect of the informational treatment was to decrease the number of STEM credits students took in the subsequent term by 0.18 credits ($p<0.1$), which represents a decrease of two percent relative to the control mean of 8.5. The second two columns estimate effects by gender. Consistent with overconfident men adjusting their relative performance beliefs downwards, the negative effect on STEM credits is driven entirely by men. Men decreased their STEM credits by 0.28

[23]The exception to using two-digit CIP code is economics (45.06), which I code separately from the rest of the social sciences (45).

[24]Fewer than two percent of control students do not appear in the data in the semester following the intervention.

[25]Treatment effects on STEM course-taking outcomes estimated without covariates are included as Appendix Table A.11. The results are very similar.

credits (three percent; $p<0.05$) while women decreased theirs by a statistically insignificant 0.079 (one percent). I cannot reject that men's and women's behavior change equally. The gender gap in STEM credits absent the intervention is two credits, so the treatment shrinks the gap by roughly ten percent.

I find a small, marginally significant average effect on the extensive margin of STEM: a decrease in the likelihood of taking any STEM courses by 1.4 percentage points (1.5 percent; $p<0.1$). The points estimates for men and women are identical to three digits and statistically indistinguishable. For high-performing students, I test for differential effects on STEM course-taking by treatment arm (Appendix Table A.12) but find none.

Taken together, the estimated effects of the informational intervention on students' beliefs and subsequent behavior imply that men's overly confident beliefs about their relative performance are partially responsible for their higher rates of STEM persistence. By inducing them to accurately revise their beliefs about their relative performance, the experiment caused men to take fewer STEM credits. Women, on the other hand, revised their beliefs in a direction that should make them less underconfident about their relative performance, but did not change their behavior. Male overconfidence rather than female underconfidence appears to be a determinant of the gender gap in field specialization.

As a robustness check, I estimate treatment effects on STEM course-taking outcomes but limit my sample to students who responded to the post-intervention survey. The results, shown in Appendix Table A.13, produce very similar point estimates. As an additional robustness check, I re-estimate treatment effects on relative performance beliefs, adjusting for survey response using inverse probability weights that reflect how likely a student is to respond to the survey based on their observable characteristics. In this exercise, survey respondents who closely resemble non-respondents are given more weight. The results are included as Appendix Table A.14. The point estimates are similar to the ones in Table 1.3. Both exercises confirm that differential survey response is not leading to a spurious conclusion about the relationship between changes to beliefs and changes to behavior.

A natural question arising from the negative effect on STEM course-taking for male students is which types of courses they took instead. As an exploratory analysis, I test for effects on credits taken in other subjects, which I separate out by non-economics social science, psychology, business and public policy, humanities and the arts, and all other subjects. The results, included as Appendix Table A.15, indicate that the decrease in STEM credits for men may have corresponded to a shift into psychology, humanities and arts, and other courses, but the effects are not statistically significant.[26]

[26]I also investigate whether the intervention changed the difficulty of courses students take by estimating effects on an average course difficulty outcome. I calculate the proportion of students who withdrew from a

In designing an intervention that targets students' beliefs about their ability to succeed in STEM, I ultimately am interested in their choice of college major. Because the studied students are still early in their academic careers, this outcome does not yet exist. I have so far focused on course-taking as a short-term proxy for and important precursor to major choice. I also use additional outcomes and the effects on course-taking to speculate on major choice.

I pre-specified two outcomes capturing students' subjective intent to major in STEM and their interest in the field, both based on survey responses. The first is simply whether they stated in the post-intervention survey that they planned to major in a STEM subject. The second is an index aggregating stated intentions and interests, which I refer to as a STEM interest index. It combines items about their general interest in STEM, their intention to seek academic advising in a STEM field, and their intention to take subsequent STEM courses.[27] I find small, negative, statistically insignificant effects on subjective STEM intent and small negative effects on STEM interest (included in Appendix Table A.16).

As a complement to these pre-specified analyses, I estimate treatment effects on students' predicted STEM degree receipt. The basic idea is straightforward and intuitive, and follows Athey et al. (2019). A prior cohort of students serves as the basis for predicting STEM degree receipt as a function of a set of demographic and academic characteristics, including the courses they take in all possible subjects. I save the estimated parameters from this prediction and apply them to the experimental sample to get their predicted probability of majoring in STEM. I can then estimate treatment effects on this predicted probability. This provides a sense of how substantively important the short-term treatment effects are and, with some assumptions, this provides an unbiased estimate of the ATE on the long-term outcome.[28] The bottom panel of Appendix Table A.16 shows estimates for treatment effects on predicted long-term degree. The estimated effects for all students as well as for men and

course in the three previous academic years, then take the average of that proportion over the courses students took in the semester following the intervention. I find a very small negative but statistically insignificant effect for men (not shown). It's possible that the treatment shifted men into easier courses, but the evidence is weak.

[27]The index is constructed following Kling et al. (2007), where I standardize each variable using the control group mean and standard deviation, impute missing values (for individuals with at least one valid index component) with the treatment-assignment group mean, and then take the unweighted mean across the standardized, imputed components.

[28]Along with a standard unconfoundedness assumption, the two additional assumptions required in order to get an unbiased treatment effect are as follows. (1) Surrogacy: the long-term outcome is independent of the treatment conditional on the full set of surrogates (i.e., pre-treatment X's and short-term outcomes). In my case, this means the treatment affects STEM majoring only through observed student characteristics and accumulated credits and not through any other channel. (2) Comparability: the conditional distribution of the primary outcome conditional on the surrogates is the same in the two samples. This would be violated if the relationship between course-taking and major choice changed over time, or if the treatment somehow changed the relationship.

women are small, negative, and not statistically significantly different from zero.

Though not strong evidence, these findings are consistent with men being discouraged by the intervention. However, the magnitudes imply that any effects of the intervention on longer-term STEM persistence and major choice are likely to be small.

1.6.3 Mechanisms

Much of the prior research on feedback provision, in academic and other settings, has focused on effort and performance as an outcome (Ashraf et al., 2014; Azmat et al., 2019; Azmat and Iriberri, 2010; Bandiera et al., 2015; Dobrescu et al., 2019; Goulas and Megalokonomou, 2015; Tran and Zeckhauser, 2012). Understanding how students adjust their effort in response to feedback is interesting in its own right, as educators care about improving performance, and could also be an important mechanism through which the intervention changes students' behavior. Students who received a negative shock to their beliefs might decrease their effort due to a discouragement effect; on the other hand, they might increase their effort if they realize their performance is not adequate for a STEM major.

I pre-specified two effort and performance measures as secondary outcomes: students' score on the final exam, and their final grade in the course.[29] I estimate treatment effects on final exam and final course scores, both measured as percent scores out of 100 (included as Table 1.6). There is no evidence that the intervention affected performance for men, women, or students as a whole. Although the point estimates for both final exam and final course performance are negative for men (-0.013 and -0.141, respectively), the lower bounds of the 95 percent confidence intervals imply that men could have at most decreased their final exam and course performance by less than a percentage point, suggesting effort and performance were not a key mechanism through which changing beliefs affected behavior.

The intervention could change students' beliefs about their ability to succeed in STEM, which could serve as an intermediate channel between their beliefs about their performance and their behavior. To measure this, I construct an index capturing students' beliefs about their ability to succeed in STEM, which aggregates responses to items about their grades being "good enough" for STEM, a series of STEM-self-efficacy items, and items about identifying with being a "math person" or "science person". Like with the STEM interest index, the construction of the success index follows Kling et al. (2007). The results are included as the last panel of Table 1.6. The effects of the intervention on this success

[29] One course, EECS 183, had a final project in lieu of an exam, so I use scores on that for the final exam measure. One section of the economics course allows students to opt out of the final exam (they can drop their lowest score, so many choose not to take the final), so I do not include it in my analyses of final exam performance.

index are small and insignificant: positive 0.013 standard deviations for men, 0.035 standard deviations for women, and no detectable difference by gender.

There are theoretical reasons to expect that certain types of students' beliefs and behavior would be particularly responsive to an informational intervention. To further explore mechanisms, I report treatment effect heterogeneity along several additional pre-specified and exploratory dimensions.

There is a strong theoretical reason to believe that the informational intervention would operate differently depending on a student's pre-intervention beliefs. We would expect those who began the semester relatively underconfident to update their beliefs and behavior in a positive direction, while those initially overconfident should do the opposite. To test this, I estimate treatment effects based on whether a student under- or over-predicted their course percentile in the pre-intervention survey, for each of five key outcomes (absolute value of percentile belief error; signed percentile error; underestimating the course median for STEM majors; overestimating the median; and number of STEM credits one semester later). Appendix Table A.17 tells a consistent story about belief updating, especially for beliefs about the STEM median. The initially underconfident students update their belief about their own percentile upwards and correct their overestimation of the STEM median. The initially overconfident students update their percentile beliefs slightly downward and correct their underestimation of the median. However, the two groups have similar estimated treatment effects on STEM credits one semester later. I do a similar exploratory exercise where I instead interact the treatment indicator with a continuous measure of the student's error at the beginning of the semester (Appendix Table A.18). These results similarly suggest that the students who are initially the most overconfident update their beliefs downward by the most (or, equivalently, that those who are initially the most underconfident update upwards more). The interaction term for the effect on STEM credits is negative (which would mean students who are initially the most overconfident respond more negatively to the information) but not statistically significant. The results are broadly consistent with a story of a reduction in relative overconfidence causing a reduction in STEM specialization.

Related to the above, we might expect students who enter the semester lacking information about college-level coursework and standards to be particularly susceptible to an informational intervention. As an exploratory analysis, I proxy a pre-treatment lack of information with student level, operationalized as first year or sophomore standing versus junior or senior, and estimate effects by level (Appendix Table A.19) Though I lack the power to make precise comparisons, the point estimates by student level suggest that students earlier in their college career change their beliefs and behavior more in response to the intervention. (Even independent of effects on beliefs, we would not expect upper year

students to change their course-taking behavior by much, since the cost of switching their field specialization is much higher.)

A student's intended major at the beginning of the course might affect how they update their beliefs and change their behavior. Inframarginal students—those not even considering a STEM field—might be less moved by the intervention, while those considering a STEM major may find the information more salient and react more. Appendix Table A.20 shows treatment effects on the same five outcomes as above, by whether students indicated in the pre-intervention survey that they planned to major in a STEM subject. Although I cannot reject equality of treatment effects by intended major for all outcomes, the results suggest that it is students already interested in STEM who change their beliefs and behavior more.

Similarly, I test for heterogeneity in effect by whether a student had declared a major at the time of the intervention (Appendix Table A.21; this analysis was not pre-specified). We would expect behavior to change more for students with lower switching costs, i.e., those who had not yet declared a major. Consistent with this hypothesis, all of the negative effect on STEM credits is due to students who had not yet declared a major by the semester of the intervention. Undeclared students (but not declared students) updated their beliefs about the STEM median, while the opposite is true of own percentile beliefs. This suggests that beliefs about other STEM majors are more salient for behavior.

One advantage of my setting relative to previous work is that I am able to study students in multiple STEM fields. Although the phrase "STEM" is often used to refer to fields with similar characteristics, there is considerable variation in key factors such as the proportion of women and mathematical intensity. The seven courses in my study vary in these ways as well as in course content, grading structure, and more. I report estimated treatment effects by subject (shown in Appendix Table A.22). I estimate these using a single regression with subject-by-treatment interactions. I also test for joint significance of the subject interactions. I find mixed evidence that the treatment effect varied by course. There is some evidence that students' beliefs and behavior changed the most in the subjects where they were previously the most incorrect, but overall I lack the power needed to make precise comparisons across subject.

As an exploratory dimension, I estimate heterogeneity by the gender composition of the course, to see if students respond differently in more male-dominated fields. The results, in Appendix Table A.23, suggest that students correct relative overconfidence more in subjects that are more heavily male, and men in more male-dominated courses may respond more negatively in their STEM course-taking than men in more female fields.[30] This would be

[30]From most to least male-dominated, the proportion male is: Physics (73 percent men), Engineering (70 percent), Computer Science (62 percent), Economics (54 percent), Statistics (47 percent); Chemistry (47

consistent with men being more biased in more male-dominated fields, possibly because of gender stereotypes, and therefore being more susceptible to information.

Though I generally lack the statistical power to make comparisons across subgroups, I interpret these heterogeneity results as consistent with a world where students update their beliefs in the direction of the truth, and where students who we would expect to be on the margin of specializing in STEM (e.g., younger students, undeclared students, and students already interested in STEM) change their behavior the most.

1.7 Discussion

This work lies at the intersection of two canonical economic frameworks. The first is a discrete choice model of field specialization, first formalized by Roy (1951). In the Roy model and more recent variants (Altonji, 1993; Altonji et al., 2016; Arcidiacono, 2004; Arcidiacono et al., 2016), individuals choose a field that maximizes their expected utility. Beliefs about the individual's field-specific ability are an input into the expected value of that field; all else equal, students with higher beliefs about their ability in STEM are more likely to choose STEM. The second framework is one of Bayesian updating and learning over time (e.g., Mobius et al., 2014; Coffman et al., 2019). In this framework, individuals observe their true ability with noise, and as they receive additional signals in the form of academic performance and other feedback, they update their beliefs in the direction of the truth.

An implication of these models is that, assuming there is a positive relationship between beliefs about major-specific ability and the expected payoff to a major, those who are performing better in STEM than they expected should be (weakly) more likely to pursue STEM, while those who receive a negative signal should be (weakly) less likely. If men are particularly overconfident and women are particularly underconfident about their performance in STEM, receiving information should lead fewer men and more women to persist in the field. Furthermore, we would expect the largest changes for those who receive the largest information shock, i.e. those who are the most under- or overconfident at baseline. However, even a large shock to beliefs about ability may not be sufficient to change behavior if a student is far from the margin due to strong underlying taste (or distaste) for STEM, strong non-STEM ability, or if frictions such as stereotypes or confirmation bias prevent them from incorporating the information.

Consistent with the belief updating framework, I find that students do correctly revise their beliefs when provided with information. Both men and women correct their beliefs about how other STEM majors perform. Men but not women correct their beliefs about their

percent), and Biology (35 percent).

own relative course rank. This somewhat mixed finding is part of a somewhat mixed prior literature. Although some studies have found that women tend to update more conservatively than men (Buser et al., 2018; Mobius et al., 2014; Coutts, 2019) and that people update less when the information is about a gender-incongruent domain (Coffman et al., 2019), others find the opposite (Goulas and Megalokonomou, 2015; Owen, 2010).

A natural question arising from the observed gender differences in beliefs—absent the intervention—is how those beliefs are formed and why they persist. One possibility is that students are incorporating signals from other sources like standardized test scores and STEM courses they took previously, and men have received signals that are more positive than women. I can investigate this in the data, and while men are more likely to have taken calculus in high school and have higher quantitative test scores, controlling for all of these factors does not change the gender gap in beliefs. Theory paired with lab-based studies of belief updating suggest that exaggerated stereotypes about groups (e.g., men are much better at quantitative subjects) can persist despite very small true differences, due to people using mental shortcuts to make predictions about themselves or others (Bordalo et al., 2016). This would explain men overestimating and women underestimating their own quantitative ability.

Consistent with field-specific beliefs mattering for specialization, men updating relative beliefs downwards leads to them taking fewer STEM credits. Though women update in a way suggesting an increase in their relative performance beliefs, their behavior does not change. Understanding why women's choices are unmoved is critical to fully understanding gender differences in field choice. This could be explained by women having a comparative advantage in non-STEM, which remains even after revising STEM beliefs (Breda and Napp, 2019). Gender differences in STEM and non-STEM performance support this: although control men and women in the sample have indistinguishable GPAs in their college STEM courses, women do significantly better in non-STEM subjects. It could also be the case that factors other than academic beliefs matter most for women. Using survey data to estimate a structural model, Zafar (2013) finds that gender differences in preferences and tastes, rather than confidence about academic ability, explain the gap in major choice. Recent interventions by Porter and Serra (2019), Li (2018) and Bayer et al. (2019) also suggest that factors such as information about and interest in the field and the presence of female role models can affect women's choices. Finally, it could be true that while women care about their performance, their *relative* rank or their performance compared to other STEM majors is less salient than it is for men. This hypothesis is supported by research finding that men have stronger preferences for competitive environments and respond more to information about the competition they face (Niederle and Vesterlund, 2011; Buser et al., 2014; Berlin

and Dargnies, 2016). Because women's beliefs about their own relative rank do not change in response to the intervention, I cannot rule out that their behavior would change if they updated those beliefs rather than or in addition to their beliefs about the typical STEM student—though changing those beliefs may be difficult.

1.8 Conclusion

The topic of gender differences in college field specialization and its implications for the labor market is one of great interest to educators and other policymakers. There is a strong theoretical and empirical basis for believing that gender differences in students' perceptions of relative performance in STEM may be contributing to gender gaps in college major choice, but the causal evidence identifying this mechanism has thus far been limited. To understand this mechanism, I ran a field experiment across seven large introductory STEM courses at a selective university. My primary treatment entailed providing students with information about their performance relative to their classmates and relative to STEM majors. I combine survey data on students' beliefs with administrative data on academic behavior to investigate behavioral changes and the mechanisms behind them.

Consistent with prior empirical findings about gender differences in beliefs, I find that men, particularly the lowest performing ones, are substantially more overconfident than women about their relative performance in STEM courses. Consistent with theoretical work that beliefs matter for educational choices, I find that providing information helps correct this overconfidence and close gender gaps in STEM persistence, with overconfident men updating their beliefs and adjusting their STEM course-taking downward. While the direction of the changes is perhaps surprising, these findings advance our understanding of how beliefs factor into academic decisions. Prior work has disagreed on whether female underconfidence rather than male overconfidence should be targeted to close gender gaps, but my work supports the latter. This conclusion is consistent with several recent papers that use observational data to argue that much of the gender gap in STEM is due to lower-achieving men persisting despite their marginal qualifications (Bordón et al., 2020; Cimpian et al., 2020).

I cannot yet observe how the short-term changes to beliefs and behavior induced by the informational intervention translate to longer-term, consequential decisions such as major declaration and degree receipt. The passage of time and follow-up data will reveal whether information provision permanently discouraged men from STEM and shrank gender gaps.

While a full welfare analysis is beyond the scope of this study, a number of factors should be weighed in evaluating the effects of an informational intervention. It will be important to see whether the intervention simply shifted the timing of men leaving STEM,

rather than discouraging those who would have otherwise stayed; the former implies welfare improvements for men who figure out their comparative advantage sooner as a result of the intervention. On the other hand, if the information provision discouraged men who would have otherwise persisted in STEM, whether they are better off will depend on the major they choose instead and the associated labor market and non-pecuniary outcomes. Low-performing men leaving STEM could also have several important spillover effects on the students who remain. Some majors have capacity constraints which may be eased by having fewer students, freeing up spots for higher-achieving students and women. The changing composition of students in STEM courses to be less male and less low-achieving may also have peer effects on remaining students.

This study provides the first experimental evidence that gender differences in students' beliefs about their relative performance—male overconfidence in particular—contribute to gender gaps in STEM, but several important questions remain unanswered and are ripe for future research. This paper studied only students in STEM classes, who had already shown a high level of interest in STEM, and focused on STEM-specific beliefs. In future work, it will be important to study students' beliefs about their performance in non-STEM subjects, where gender differences may be less stark or even reversed. Likewise, non-STEM students may be even more biased about STEM than STEM students, and susceptible to interventions encouraging STEM. Understanding the full set of students' beliefs about who pursues various fields and their own field-specific potential is critical for understanding field specialization decisions.

While I included students studying multiple STEM subjects, this single study lacks the statistical power to precisely compare across STEM fields. We might expect biology—a predominantly female field—to show different patterns in students' beliefs and different responses to intervention than a male-dominated field like engineering. Future work should explore this further. Finally, this paper studies students at a single, highly selective institution, the University of Michigan. It is possible that the degree of overconfidence among the students in my sample is related to their backgrounds and high levels of prior achievement; different populations of students may hold very different beliefs about relative performance and react differently to information.

Although the magnitudes of my treatment effects are modest, they are the result of an extremely light-touch, low-cost intervention—a single tailored email that can easily be sent to a large number of students. A more intensive or repeated intervention may be effective at changing beliefs and behavior even more. Taken in context, my findings suggest that biased beliefs about relative academic performance are one important piece of the large, complex issue of decisions about field specialization and gender differences in STEM.

Figure 1.1: Experimental Design

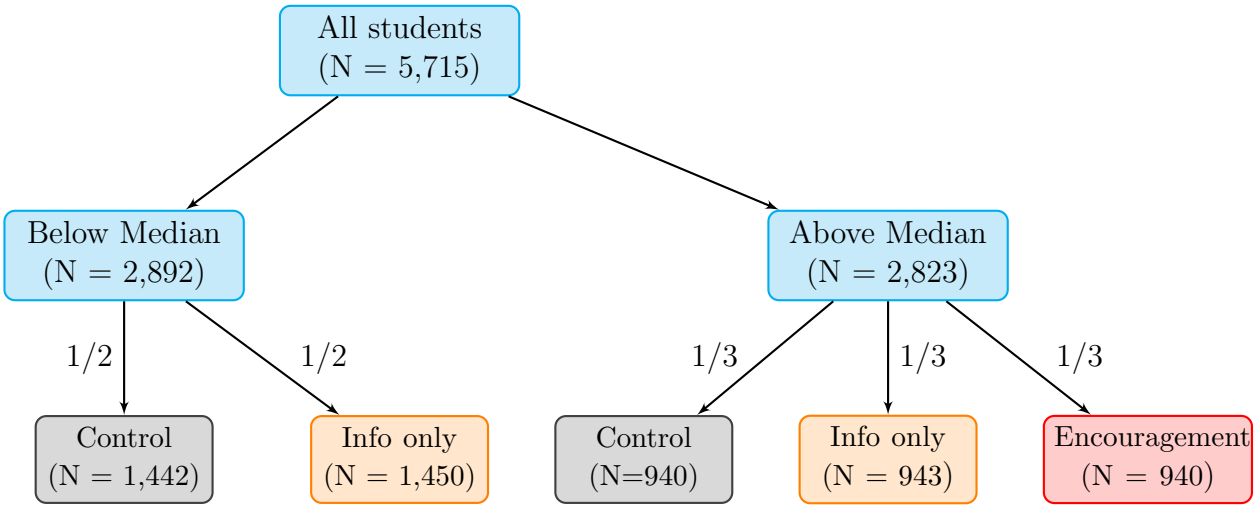
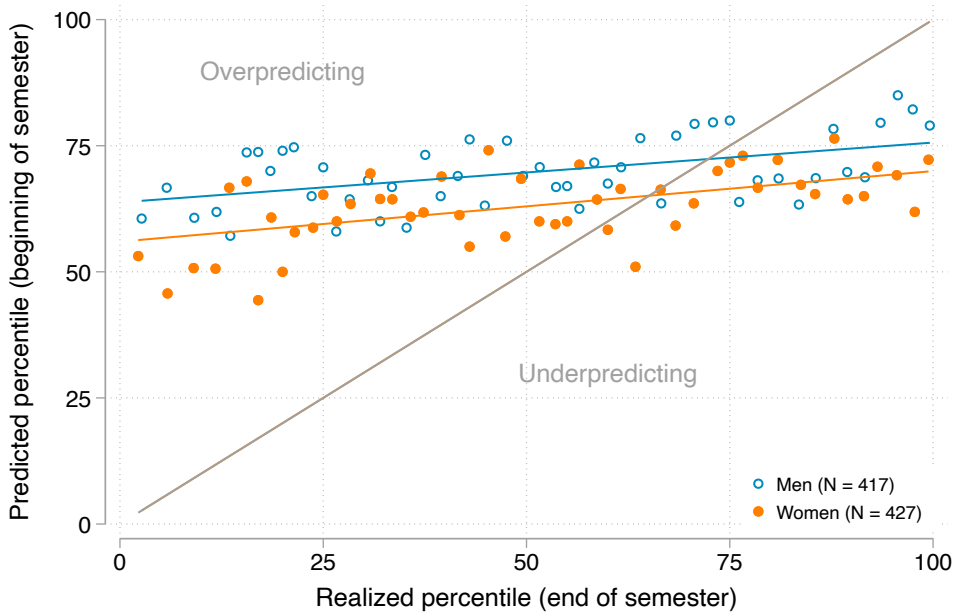
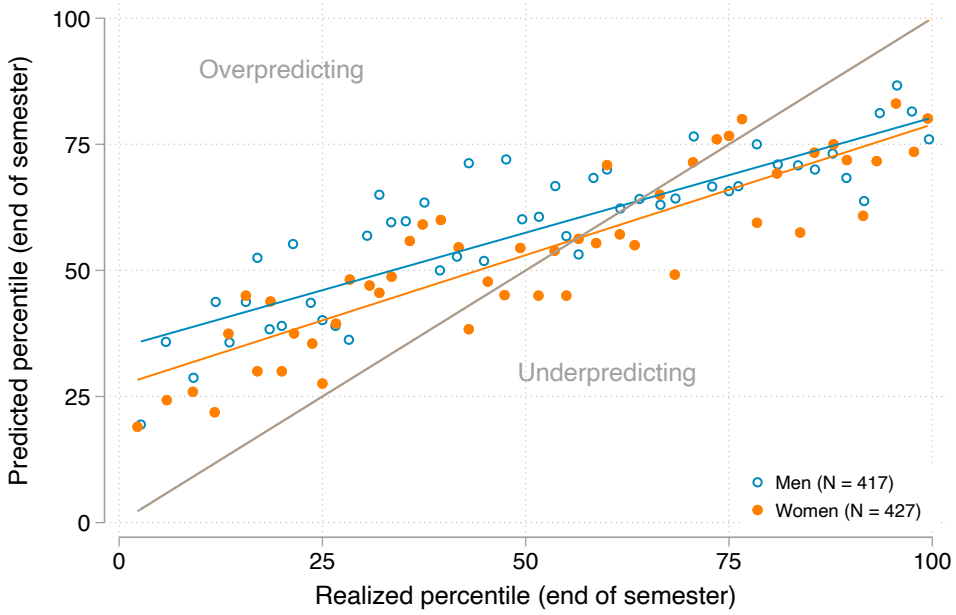


Figure 1.2: Control Student Beliefs about Own Percentile by Gender



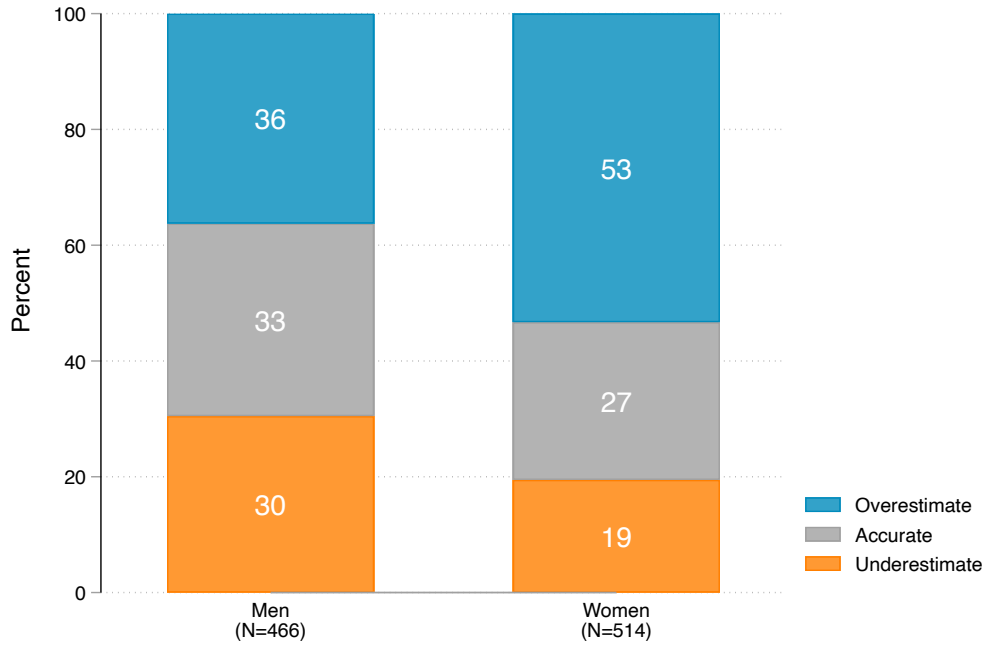
(a) Beginning of Semester Beliefs



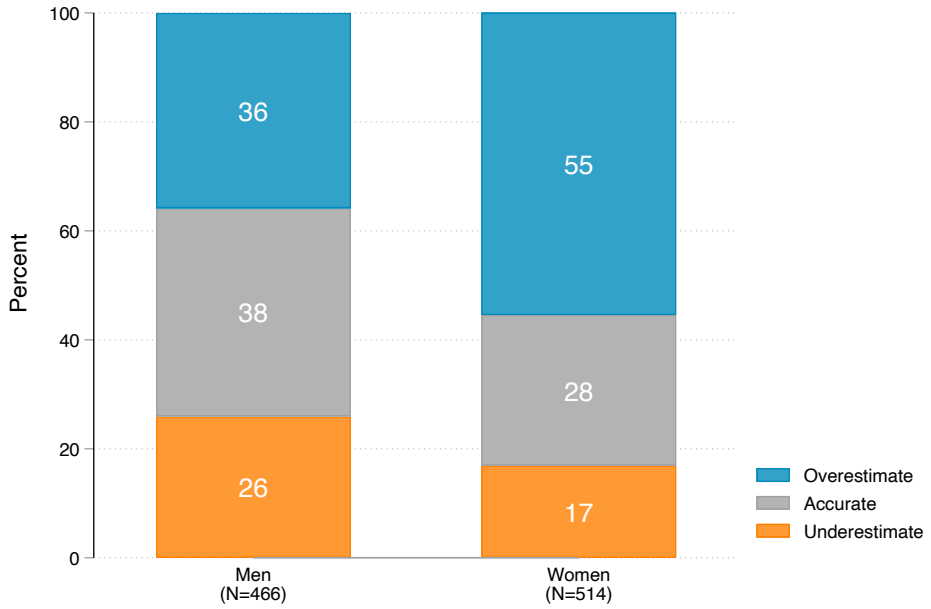
(b) End of Semester Beliefs

Notes: Sample restricted to control students who responded to the question about percentile beliefs on both the pre- and post-intervention surveys. X-axis measures students' realized percentile within the course, measured at the end of the semester. Y-axis measures what students predict their final percentile will be when asked on the survey. Figure is a binned scatterplot plotting the average values within 50 equally-sized bins of students.

Figure 1.3: Control Student Beliefs about Course Median for STEM Majors by Gender



(a) Beginning of Semester Beliefs



(b) End of Semester Beliefs

Notes: Sample restricted to control students who responded to the question about the median on both the pre- and post-intervention surveys. Overestimating means the student thinks the median is higher than it is (e.g., they median is a B and they think it is a B+), while underestimating means they think the median is lower than it is.

Table 1.1: Balance by Assignment to Treatment, Full Sample

	Control mean	Treatment mean	p-value	N non missing
Female	0.479	0.474		5,715
Class standing (omitted: senior)				
First year	0.433	0.417	0.317	5,715
Sophomore	0.387	0.403	0.551	
Junior	0.132	0.132	0.819	
Race/ethnicity (omitted: American Indian or multiple race/ethnicities)				
White	0.558	0.543	0.262	5,554
Hispanic	0.070	0.068	0.423	
Asian	0.254	0.289	0.159	
Black	0.038	0.025	0.200	
Declared major (omitted: other)				
Undeclared	0.560	0.559	0.606	5,715
Engineering	0.232	0.236	0.485	
Math, science, or economics	0.095	0.094	0.658	
Academic and demographic characteristics				
In-state	0.524	0.520	0.363	5,715
Prior college GPA	3.38	3.43	0.662	2,385
Math placement score (std.)	-0.080	0.057	0.432	5,478
ACT English	32.3	32.6	0.885	3,151
ACT Math	30.9	31.3	0.990	3,151
ACT Reading	32.0	31.8	0.006	3,151
ACT Science	30.9	31.1	0.297	3,151
SAT Math	705	714	0.556	3,407
SAT Verbal	642	647	0.875	3,407
HS GPA	3.88	3.89	0.546	4,952
Took calculus in HS	0.814	0.838	0.427	5,104
Max parental education (omitted: less than high school)				
High school	0.071	0.070	0.271	5,641
Some college	0.064	0.051	0.403	
Bachelor's	0.253	0.241	0.431	
Grad or professional degree	0.588	0.617	0.603	
Family Income (omitted: less than $50,000)				
$50,000-100,000	0.182	0.189	0.213	4,374
Above $100,000	0.625	0.643	0.542	
P-value on F-test of all X's		0.840		5,715
Total N	2,382	3,333	5,715	

Notes: "Treatment" includes students receiving either treatment arm. P-values based on a regression of the characteristic on treatment status, controlling for strata. I also test for differences in missingness rates on all variables and find none. F-test tests for joint significance of all listed characteristics (except for female, which is blocked on) as well as missingness rates in predicting treatment, controlling for strata.

Table 1.2: Decomposition of Gender Gap in STEM Credits by Relative Performance Beliefs and Other Covariate Components (Control Students Only)

Female - male gap in STEM credits	-2.15		
	(0.28)		
Covariate	Gap explained by covariate	Percent of total gap	Percent of explained gap
Own percentile belief	-0.04	2%	4%
	(0.04)		
STEM median belief	-0.11	5%	10%
	(0.05)		
Realized percentile	-0.02	1%	2%
	(0.02)		
Demographics	0.02	-1%	-2%
	(0.05)		
High school achievement	-0.02	1%	2%
	(0.10)		
Math placement score	-0.15	7%	14%
	(0.06)		
Prior college achievement	-0.04	2%	4%
	(0.05)		
Student level	0.00	0%	0%
	(0.03)		
Declared major	-0.69	32%	66%
	(0.16)		
Total explained	-1.05	49%	100%
Total unexplained	-1.10	51%	-
N	918		

Notes: Decomposition follows Gelbach (2016) and is implemented using b1x2 command in Stata. STEM credits measured in the semester following the one when students took the course. Own percentile belief is a student's 1-100 prediction of their own final course percentile, measured in the end of semester survey. STEM median belief measured as two dummy variables for whether a student is over- or underestimating the course median for STEM majors, measured in the end of semester survey. Demographics include race, parent education, family income, and in-state status. High school achievement includes ACT and SAT scores, high school GPA, and a high school calculus indicator. College achievement measured as prior GPA at UM. Sample limited to control students who answered both surveys.

Table 1.3: Estimated Effect of Intervention on Students' Beliefs about Themselves and Other STEM Majors, Overall and by Gender

	Absolute value error in percentile beliefs (| Predicted - realized |)			Signed error in percentile beliefs (Predicted - realized)		
	All	Men	Women	All	Men	Women
Treatment effect	-1.485**	-2.243**	-0.743	0.592	0.536	0.647
	(0.657)	(1.007)	(0.858)	(0.849)	(1.270)	(1.138)
P-value, women vs. men			0.259			0.948
Control mean	18.981	20.331	17.646	6.361	8.471	4.276
N	2,358	1,166	1,192	2,358	1,166	1,192
	Underestimating course median for STEM majors			Overestimating course median for STEM majors		
	All	Men	Women	All	Men	Women
Treatment effect	-0.033**	-0.052**	-0.016	-0.023	0.007	-0.051**
	(0.015)	(0.022)	(0.019)	(0.018)	(0.026)	(0.026)
P-value, women vs. men			0.220			0.111
Control mean	0.206	0.257	0.159	0.46	0.368	0.545
N	2,632	1,291	1,341	2,632	1,291	1,341

Notes: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$. Treatment effects for all students estimated from a regression of the outcome on assignment to either treatment, controlling for student academic and demographic characteristics and randomization strata dummies. Treatment effects by gender estimated from a single regression of the outcome on assignment to the either treatment, female, and treatment-times-female, controlling for student academic and demographic characteristics and randomization strata dummies. Robust standard errors reported. All beliefs outcomes based on response to post-intervention survey. Realized performance measured mid-semester, at the time of intervention.

Table 1.4: Estimated Effect of Intervention on Students' Beliefs, by Gender, Mid-Semester Performance, and Treatment Arm

	Signed error in percentile beliefs (Predicted - realized)			Underestimating course median for STEM majors			Overestimating course median for STEM majors		
	All	Men	Women	All	Men	Women	All	Men	Women
Below median students									
Info-only effect	-1.349	-3.735**	0.736	-0.065***	-0.088**	-0.044	0.035	0.064*	0.010
	(1.243)	(1.881)	(1.665)	(0.023)	(0.035)	(0.030)	(0.026)	(0.037)	(0.036)
P-value, women vs. men			0.077			0.345			0.293
Control mean	17.437	21.442	13.765	0.26	0.318	0.209	0.372	0.281	0.451
N	1,058	497	561	1,215	569	646	1,215	569	646
Above median students									
Pooled effect	2.375**	4.095**	0.553	-0.003	-0.020	0.014	-0.078***	-0.043	-0.115***
	(1.160)	(1.710)	(1.543)	(0.018)	(0.029)	(0.022)	(0.026)	(0.037)	(0.036)
P-value, women vs. men			0.123			0.353			0.168
Info-only effect	1.350	2.226	0.400	-0.001	-0.009	0.007	-0.077**	-0.041	-0.113***
	(1.366)	(2.037)	(1.761)	(0.021)	(0.034)	(0.026)	(0.030)	(0.043)	(0.042)
P-value, women vs. men			0.493			0.700			0.230
Info + encouragement effect	3.385***	5.347***	1.257	-0.006	-0.033	0.023	-0.073**	-0.039	-0.109***
	(1.287)	(1.842)	(1.790)	(0.021)	(0.032)	(0.026)	(0.030)	(0.043)	(0.042)
P-value, women vs. men			0.112			0.172			0.243
P-value, info vs. info+enc	0.105	0.085	0.619	0.819	0.450	0.570	0.910	0.968	0.927
Control mean	-8.111	-7.081	-9.232	0.134	0.181	0.086	0.577	0.475	0.681
N	1,300	669	631	1,417	722	695	1,417	722	695

Notes: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$. Only above-median students were eligible for the information-plus-encouragement treatment; all below-median treated students received information only. Effect of information-only treatment for below-median students and either treatment (pooled) for above-median students estimated from a regression of outcome on an indicator for receiving either treatment, an indicator for being above the course median at time of randomization, and their interaction. To estimate effects on men and women, a full three-way interaction between treatment, female, and above-median is added. Treatment effects of the information-only and info-plus-encouragement intervention for above-median students estimated only on the sample of above-median students using the same specifications as above, but with two separate treatment indicators. All regressions control for student academic and demographic characteristics and randomization strata dummies. Robust standard errors reported. All beliefs outcomes based on responses to post-intervention survey. Realized performance measured mid-semester, at the time of intervention.

Table 1.5: Estimated Effect of Intervention on Students' STEM Course-taking, Overall and by Gender

	Number of STEM credits one semester post intervention			Took any STEM courses one semester post intervention		
	All	Men	Women	All	Men	Women
Treatment effect	-0.182*	-0.276**	-0.079	-0.014*	-0.014	-0.014
	(0.095)	(0.129)	(0.140)	(0.007)	(0.009)	(0.012)
P-value, women vs. men			0.303			0.975
Control mean	8.507	9.476	7.454	0.91	0.936	0.881
N	5,715	2,993	2,722	5,715	2,993	2,722

Notes: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$. Treatment effects for all students estimated from a regression of the outcome on assignment to either treatment, controlling for student academic and demographic characteristics and randomization strata dummies. Treatment effects by gender estimated from a single regression of the outcome on assignment to the either treatment, female, and treatment-times-female, controlling for student academic and demographic characteristics and randomization strata dummies. Robust standard errors reported. Course-taking outcomes based on University of Michigan administrative data.

Table 1.6: Estimated Effect of Intervention on Students' Performance and Beliefs about Ability to Succeed in STEM, Overall and by Gender

	Final exam or project score (out of 100)			Final course score (out of 100)		
	All	Men	Women	All	Men	Women
Treatment effect	-0.167	-0.013	-0.334	0.004	-0.141	0.164
	(0.332)	(0.454)	(0.486)	(0.186)	(0.252)	(0.275)
P-value, women vs. men			0.630			0.415
Control mean	80.917	81.666	80.107	83.974	84.62	83.273
N	5,323	2,785	2,538	5,648	2,961	2,687
	STEM success index (std. dev. units)					
	All	Men	Women			
Treatment effect	0.024	0.013	0.035			
	(0.025)	(0.035)	(0.035)			
P-value, women vs. men			0.656			
Control mean	0	0.116	-0.108			
N	2,687	1,317	1,370			

Notes: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$. Treatment effects for all students estimated from a regression of the outcome on assignment to either treatment, controlling for student academic and demographic characteristics and randomization strata dummies. Treatment effects by gender estimated from a single regression of the outcome on assignment to the either treatment, female, and treatment-times-female, controlling for student academic and demographic characteristics and randomization strata dummies. Robust standard errors reported. Performance outcomes based on University of Michigan administrative data. STEM success index is based on post-intervention survey responses and aggregates items about being "good enough" for STEM, self-efficacy, and STEM identity.

CHAPTER II

Ahead of the Curve: Grade Signals, Gender, and College Major Choice

2.1 Introduction

The choice of college major is a consequential at both the individual and aggregate level. There is wide dispersion in earnings by major, with the highest-earning making substantially more than the lowest-earning over their lifetime (Webber, 2019). Since different types of students tend towards different types of majors, earnings differences by major have implications for income gaps by gender, race, and other characteristics. Furthermore, the allocation of students across majors and workers across jobs affects overall economic efficiency. If there are barriers or frictions keeping some individuals out of certain fields, then removing these barriers would lead to more efficient sorting and increase overall productivity (Hsieh et al., 2019). Finally, there is considerable policy interest in encouraging students to study certain fields, both to address shortages as well as to close gender and race gaps and address equity concerns.

Prior work has suggested that grades student receive are a key input into major choice, and that the relationship varies by gender (e.g., Avilova and Goldin, 2020). However, much of this work is observational, and it is not clear how much the relationship between grades and behavior is about the effort, ability, and motivation behind the grade versus the signal the grade itself provides for students or the hurdles it may remove (psychological or administrative). If the grade itself matters and matters differently for women versus men, educators could use grades as a policy lever to induce students into certain fields and potentially close gender gaps. Doing this effectively requires knowing if and how grades affect choices conditional on endogenous inputs, as well as if and how the effect varies by gender.

To answer these questions, I study a natural experiment within the economics department

of a large university, which changed the grading policy in its introductory economics courses to award higher grades. Specifically, the instructors went from a curve guaranteeing that 25 percent of students got a grade in the A range and 33 percent in the B range to guaranteeing at least 35 percent got some type of A and at least 75 percent got an A or B grade. This policy introduced variation in the letter grade students received, even conditional on their underlying effort and performance.

Rich administrative data provide information on students' final grades as well as their raw scores; I track their subsequent course-taking and academic decisions with full transcript data. I leverage the variation introduced by the change in the grading policy as well as naturally-occurring variation in letter grade cutoffs across semesters to compare students with the same underlying performance but who received different letter grades. This allows me to isolate the effect of the grade itself from confounding variables such as academic preparation and ability, and endogenous inputs such as interest and effort.

The results suggest that receiving one higher letter grade in introductory economics—for example, receiving a C over a C-, or an A- over a B+—increases the likelihood that a student will take the next course in the sequence by approximately 2.5 percentage points. These short term effects translate to small effects on later economics course-taking and major declaration. In my preferred specification, the effect on taking the third course in the economics sequence is a statistically insignificant 0.6 percentage points, and the effect on declaring an economics major is 0.8 percentage points. I study not just economics course-taking and major choice but also behavior across subjects. In the setting I study, introductory economics is required for admission to the selective business school, and my findings imply that relatively higher economics grades may enable students to switch to related, competitive majors like business. A student who receives a higher grade in introductory economics is 2.3 percentage points more likely to declare a business major. Students who were induced by higher grades to major in economics or business appear to have switched from other social science majors. Business is the major with the highest post-graduate earnings at the studied institution, so these changes are likely welfare-improving for the affected students. I find little evidence that women are more (or less) responsive to grades than men. The estimated effects of grades on academic outcomes are very similar and statistically indistinguishable for men versus women.

My findings suggest that grade inflation as a policy may work to retain more students—including women—within a field, but will not necessarily close gender gaps. From the perspective of an academic department or institution, the optimal policy will depend on the objective (increasing the number of economics majors versus shifting students into majors with the highest earnings) and the structure of major requirements across departments (e.g., requiring economics courses for business majors).

The paper proceeds as follows. I summarize prior related work and my contribution in Section 2.2, introduce the setting and data in Section 2.3, and describe my empirical approach in Section 2.4. I present main results in Section 2.5 and robustness checks in Section 2.6. Section 2.7 concludes.

2.2 Conceptual Framework and Prior Work on the Effect of Grades

Theoretically, there are several reasons why we would expect letter grades to causally affect students' behavior, and in particular why the effects might differ between women and men. First, many academic programs, including college majors and scholarships, set thresholds for entry or continued eligibility. Receiving, say, a C over a C- in introductory economics could mechanically allow a student to major in economics or business—which set C as the minimum grade required for entry—over a field that doesn't require economics courses. As a more extreme example, a higher grade could allow a student to retain their financial aid and make the difference between persisting and dropping out of college altogether. Students also care about their GPAs for future plans such as graduate or professional school and entering the workforce; both employers and graduate institutions use college performance to evaluate candidates. For students trying to maximize their GPAs for any of the above reasons, the grade in an introductory course helps form expectations for their GPA if they major in that subject. All else equal, a lower grade means a lower expected GPA and could nudge them towards an easier-grading major.

These explanations could play out differently by gender for several reasons. If women and men are at different points in the grade distribution and students respond to certain grades, this could result in different observed behavioral responses, even if men and women respond similarly to a given grade. For example, if more women are on the margin between a C- and C grade and receiving a C is particularly consequential, we could see a stronger response for women. In a model of comparative advantage, a student's next best major option and their grades in that subject relative to economics will determine whether a higher economics grade pushes them over the margin into economics or a related major. If women's grades (or interest) in other subjects are much higher than in economics, a higher economics grade may not be enough to shift them into economics; this would manifest as women appearing less responsive to grades. Similarly, if men's non-economics GPAs are low enough, they could be inframarginal economics majors, resulting in their behavior changing less in response to higher economics grades.

Grades could also affect behavior through their signaling value. A higher letter grade

may confer some positive utility or signal of ability to students above and beyond the other signals of their ability that they've observed. Much of the theoretical economics literature conceptualizes grades in this way, as signals by which students learn about their field-specific ability over time (Arcidiacono, 2004; Altonji et al., 2016). The existence of stereotypes about groups, such as men being relatively better at quantitative subjects, could affect how much weight is given to signals and lead to different behavioral responses by gender (Bordalo et al., 2016). There are a number of empirical studies suggesting women and men may interpret performance feedback differently, though they do not always agree on whether men versus women respond more (see, e.g., Mobius et al. (2014) compared to Goulas and Megalokonomou (2015)).

The empirical work on this topic consistently finds that grades are related to behavior, but there is a lack of strong causal evidence. A number of studies using selection-on-observables designs—i.e., comparing observationally similar students with different grades—find, unsurprisingly, that grades are associated with subsequent course-taking and major choice (Chizmar, 2000; Jensen and Owen, 2001; Rask and Bailey, 2002; Rask and Tiefenthaler, 2008; Ost, 2010; Emerson et al., 2012; Astorne-Figari and Speer, 2019; Kaganovich et al., 2020; Kugler et al., 2021). These studies come to different conclusions about whether women are more sensitive to grades than men, and whether easing grading standards in traditionally male-dominated fields could be an effective policy to close gender gaps. Rask and Bailey (2002), Rask and Tiefenthaler (2008), and Ost (2010) find support for the hypothesis that women respond more than men to grade signals in early courses. Two recent papers by Kaganovich et al. (2020) and Kugler et al. (2021) find that women are more likely to leave male-dominated STEM fields (including economics) in response to lower grades, but don't find different gender responses in other fields. Kugler et al. (2021) interpret this as women needing multiple negative signals (low grades, the presence of few other women, and stereotypes about whether a field is male or female) to leave a major at a higher rate than men. Kaganovich et al. (2020) argue that what appears to be greater grade sensitivity actually reflects a weaker underlying preference for those fields. However, Astorne-Figari and Speer (2019) and Chizmar (2000) use a similar approach and data but find no evidence that women are more likely to switch majors in response to low grades.

Studies that take a more structural approach similarly find that, consistent with a framework of learning about field-specific ability through grades, students who perform worse are more likely to switch majors (Arcidiacono, 2004; Zafar, 2011; Stinebrickner and Stinebrickner, 2014). Several of these examine differences by gender, and while Calkins (2020) suggests that women respond more to grades and improving women's grades could close gender gaps in STEM, Zafar (2013) finds that differences in preferences rather than in

beliefs about ability are responsible for most of the gender gap in major choice.

However, none of these studies have any exogenous variation in grades, and do not have finer measures of underlying performance than letter grade or overall GPA. In this study, I exploit plausibly exogenous variation in grades, and furthermore am able to much more precisely control for students' underlying effort, motivation, and preparation, to the extent they are reflected in raw course performance.

More closely related are a handful of studies exploiting plausibly exogenous variation in letter grades. Owen (2010) and Main and Ost (2014) both use regression discontinuity designs, controlling for raw score and comparing students above and below the cutoffs for letter grades in introductory economics courses. Though they use similar approaches and study similar settings, these two studies come to different conclusions. Main and Ost (2014) find no effect of receiving a higher letter grade on subsequent course-taking or major choice for any students, and no evidence of different responses by gender. Owen (2010), on the other hand, finds that receiving a higher grade in introductory economics increases the probability of majoring in economics for women but not for men. Though these RD designs are conceptually similar to my approach, they rely on a different identifying assumption and exploit a different source of variation. The validity of the RD requires that within a course, students who end up with slightly different raw scores and therefore grades are not different in other ways that could affect their outcomes. If a student has a slightly higher score as a result of a targeted effort to achieve their desired grade, or successful advocacy for a re-grade, both of which may reflect a higher interest in the subject, this assumption would be violated. My approach, on the other hand, compares students across courses and semesters rather than within, and uses an external change to the way grades are assigned. Crucially, any manipulation in underlying scores is already accounted for in the measure of raw score that I use, and does not threaten my identification strategy.

Finally, Butcher et al. (2014) use a type of policy variation similar to the current study to examine the effect of grade inflation on major choice. They exploit an anti-grade inflation policy that affected different academic departments differently. They compare previously lenient-grading departments to harder-grading departments unaffected by the new rule (including economics) and find that the policy decreased the rate of students enrolling and majoring in departments that saw grade deflation relative to those that didn't. However, the setting of that paper—Wellesley College, an all-women's institution—does not allow the authors to say anything about differential response by gender. The current paper provides the most convincing causal evidence to date on the effect of letter grades on students' academic choices, and how those effects vary by gender.

2.3 Setting, Policy Background, and Data

I study a large, selective, public flagship university in the Midwest, which I will refer to as Midwestern University or MU. In the fall of 2016, the economics department at MU changed the grading curve in its introductory courses to give out more grades in the A and B ranges. Prior to the change, instructors in the Principles of Economics courses guaranteed that 25 percent of students received a grade in the A range (A-, A, or A+) and 33 percent received a grade in the B range. After the change, at least 35 percent of students were guaranteed some type of A and at least 75 percent an A or B. This change reflected concern among economics faculty that the department was not keeping up with grade inflation across the university, and that the harsher grading was deterring students from enrolling and persisting in its courses.[1] This change was not an official one voted on by the full department; rather, the instructors teaching Principles collectively agreed to give out higher grades. There was no enforcement from above and there were no official sanctions for not complying. At MU, economics instructors have considerable independence in teaching their sections. They write their own assignments and exams, weight assignments how they like, and are not required to use the same textbook. Although the Principles instructors agreed to this new common grading curve, they had discretion over their own students' grades and implementation of the policy. There was no official announcement about the policy change, and, according to instructors, students would not have known about it when registering for fall 2016 classes. Some of the instructors announced it to their students at the beginning of the fall 2016 term, while others did not bring it to students' attention.

Though I focus on Principles I - Microeconomics, the first course in the economics sequence, the policy also extended to Principles II - Macroeconomics, the second course in the sequence.[2] Instructors in Principles II adjusted their curve one semester later than Principles I (spring 2017 vs. fall 2016) so that the change occurred starting with a single cohort. In this sense, the policy can be thought of as inflating the grades for both of the first two core courses in the department and the major.

I combine two sources of administrative data to leverage this policy change and study the effect of grades on student behavior. Learning management system (LMS) data allow me to measure student performance in Principles I. Crucially, LMS data contain students' continuous raw scores on assignments and in the course overall before letter

[1] Source: internal department memo comparing undergraduate grading policies in economics to other fields.

[2] Principles I is an advisory prerequisite for Principles II, meaning it is highly recommended but not strictly required. In practice, only around one percent of students take Principles II without having first taken Principles I.

grades are assigned. I merge these data with university student record data, which include individual-level academic and demographic characteristics (standardized test scores, high school GPA, gender, race, parental education, family income, etc.), as well as full longitudinal academic transcripts (official letter grades, courses taken, and declared major).

Since each instructor manages their own LMS page, the structure of LMS data varies across instructors and sections. Between 2014 and 2016, MU transitioned from one LMS to another, so the structure also varies over time. My empirical approach requires a measure of students' final total score in Principles I. In some cases, instructors entered a final score into the LMS, and little additional cleaning was required. In other cases, I constructed final scores based on individual assignment scores and the weighting of assignments detailed in course syllabuses. There is likely a non-trivial amount of measurement error arising from this process. I discuss this issue more in Section 2.6.2, and show that it is not substantively altering my conclusions.

Because I am studying the effect of letter grades, I limit the sample to students who complete the class and receive a letter grade, which excludes students who elect to take the course Pass/Fail. I also restrict to students with observations in the LMS data. If a student repeated the course, I use their first observation. The final sample includes 11,836 students, covering students who took Principles of Economics I for the first time between spring 2013 and spring 2018 (inclusive).[3] The dataset includes eight unique instructors, 11 academic terms, and 49 lecture sections.[4] Of the eight instructors, five taught both before and after the grading curve changed.

2.4 Empirical Strategy

My empirical strategy compares students with the same underlying economics course performance (measured as their final percentile rank within their instructor's section of Principles I) and observable characteristics, but who receive different letter final grades. By holding underlying performance constant, this strategy controls for all of the inputs that determine grades—such as effort, motivation, and prior academic preparation—and could also affect subsequent academic outcomes.

This approach uses two types of plausibly exogenous variation in letter grades. The first is variation introduced by the policy change, which increased a student's expected letter grade conditional on their raw score or percentile. Consider a simple example where an instructor

[3]I include fall and spring courses only. MU does offer Principles I during its summer term, but the courses have much smaller enrollments and are structured differently.

[4]This is not the universe of Principles I course offerings during this time, which included 61 lecture sections. A handful of instructors did not have archived LMS data available.

strictly implemented the curve. Under the old grading regime 25 percent of students were guaranteed an A grade, while under the new regime at least 35 percent were. Comparing two students who performed at the 74th percentile (right below the old cutoff), the student who took the course before the change would receive a B+, while the student who took the course after would receive (at least) an A-.

The second source of variation is naturally occurring variation in letter grades across semesters under the same official grading curve policy. Because the instructors have ultimate discretion in assigning letter grades and cannot perfectly control the composition of students in their courses or the difficulty of their exams and assignments, the same instructor may give out more A's in one semester than another. The new grading policy, which states that *at least* 35 percent of students receive A's, explicitly allows for different grade distributions. Under both policies, there is considerable variation in how many A's and B's instructors awarded. (I show evidence of this in Section 2.5.2.)

Formally, I estimate the following equation:

$$\begin{aligned} Y_{ijt} = \beta_0 + \beta_1 Grade_{ijt} + \beta_2 Percentile_{ijt} + \beta_3 Percentile_{ijt}^2 + \beta_4 Percentile_{ijt}^3 \\ + \sum_j \alpha_j Instructor_j + \delta Fall_t + \lambda Year_t + \boldsymbol{\gamma} \boldsymbol{X_i'} + \epsilon_{ijt} \end{aligned} \tag{2.1}$$

where i indexes students, j instructors, and t time periods (academic terms). The term β_1 is the estimand of interest and represents the effect of receiving a higher letter grade in Principles of Economics I. In my main specification, I estimate a constant linear effect of a higher grade; a one unit increase in letter grade is equivalent to receiving an A- over a B+, or a C over a C-. I also estimate a version of Equation 2.1 where I replace the *Grade* term with indicators for each letter grade to separately identify the effect of an A, A-, etc. Since I largely lack the power to compare effects by each grade, I prefer the linear specification.

In controlling for underlying performance, I calculate percentiles within academic term and lecture section, since this is the level at which grades are curved. (Note I calculate percentiles before excluding any students from the sample, such as those who took the course pass/fail or were taking it for a second time.) I control for a cubic in percentile to allow for a flexible relationship between percentile and the outcome.

The vector $\boldsymbol{\alpha}$ represents instructor fixed effects, and δ captures seasonality effects (the absolute performance thresholds tend to be higher in spring terms, when more engineering majors take Principles I). I include a linear time trend λ (where $Year$ denotes academic year) to allow for upward (or downward) trends in the outcomes. The vector $\boldsymbol{X_i}$ includes student gender, race/ethnicity (indicators for Black; Hispanic; Asian; Native American, Native Hawaiian, or other Pacific Islander; and multiple races), class standing (indicators for

second, third, and fourth and higher year), family income (indicators for $25,000-$49,999, $50,000-$74,999, $75,000-$99,999, and $100,000 and above), parent education (indicators for high school, some college, bachelor's degree, and graduate degree), high school GPA, an indicator for taking calculus in high school, SAT and ACT subscore percentiles, and score on the university's math placement test. Some of these characteristics are self-reported or not collected for all students, so I also includes missingness indicators for background characteristics with any missing values.

I estimate the effect of receiving a higher letter grade in Principles of Economics I on three measures of persistence within economics: indicators for taking the second course in the sequence (Principles of Economics II - Macroeconomics), taking the third course (Intermediate Microeconomic Theory), and declaring an economics major; I measure all of these within two years of completing Principles I. The outcomes are all indicators for *ever* doing the outcome. For course-taking, this simply means the student took the course at some point in the two years following Principles I. For major choice, a student gets counted as an economics major as long as they appear as an economics major in any of the subsequent semesters (even if they double major or later switch to a different major). I also study effects on major choice beyond economics, by measuring declaration of a business major, a STEM major, or a non-economics social science major. I classify subjects using two-digit Classification of Instructional Program (CIP) codes, developed and maintained by the U.S. Department of Education's National Center for Education Statistics.[5] All effects are estimated with a linear probability model. I report robust standard errors calculated with the sandwich estimator of variance.

2.4.1 Possible Threats to Identification

The identifying assumption required to interpret β_1 causally is that conditional on instructor, observable characteristics, and percentile rank in the course, the final letter grade is orthogonal to the error term. The thought experiment takes two observably similar students who take Principles I with the same instructor at the same time of year (fall or spring) and perform equally well in the class, and assigns one a higher letter grade than the other (e.g, an A- rather than a B+). The primary research question is whether receiving the higher grade makes students more likely to persist in economics or changes their academic trajectory in some way.

[5] STEM includes natural resources and conservation (CIP code 03), computer and information sciences (11) engineering (14), biological and biomedical sciences (26), mathematics and statistics (27), and physical sciences (40). Social sciences (CIP code 45, excluding 45.06 - Economics) includes anthropology, political science, and sociology. Business majors (CIP code 52) include business administration and organizational studies. Throughout Chapter II, STEM does *not* include economics.

This assumption would be violated if letter grades are not exogenous conditional on performance. For example, if certain students are able to advocate for higher grades, this could result in students with the same performance being assigned different grades. While re-grades do happen, this is only an issue for my identification strategy if the instructors change the final grade but not the underlying score. From conversations with instructors, this is exceedingly rare. Any changes tend to happen at the individual assignment level, and are entered in gradebooks. Furthermore, instructors require students to go through an appeals process and rarely if ever grant end-of-semester requests to bump up a grade close to the margin. To test for this type of selection into letter grades more formally, I examine whether, conditional on underlying performance, instructor, academic year, and time of year, letter grade predicts observable characteristics such as gender, race, family background, and academic preparation. While I control for all of these observable characteristics in my analyses, significant effects on these falsification tests could indicate differences in unobservable characteristics that could be determining both grades and outcomes. While I do find some differences (see Section 2.6.4), they are substantively small and unlikely to be responsible for effects of the magnitude that I find.

Another possible threat to identification would be another policy change contemporaneous to the grade curve change which could also affect persistence in economics and major choice. There was a substantial change to the admissions policy of the institution's business school around the same time, which could be confounding the effect of the grade policy change. I discuss this in more detail in Section 2.6.3 and argue that it is not substantively biasing my results.

2.5 Results

2.5.1 Descriptive Sample Statistics

Table 2.1 presents mean characteristics for the sample, both overall and by gender. 39 percent of students who took Principles of Economics I during the sample period are women. Nearly two thirds of students are White, and another quarter are Asian. Very few underrepresented minority students are in the sample: 3 percent are Black, 5 percent are Hispanic, less than 1 percent are Native American, Native Hawaiian, or other Pacific Islander, and 3 percent are more than one race. The majority (77 percent) of students who take Principles I do so for the first time in their first year of college, while another 18 percent took the course in their second year; fewer than 6 percent took the course in their third year or later.

Students at MU come from very socioeconomically advantaged backgrounds. The

majority of students (53 percent) have family income in the highest category of $100,000 and above; conditional on having a reported family income, over two thirds are in this category.[6] Similarly, 58 percent of students have a parent with a graduate or professional degree, and only 8 percent are first-generation college students (meaning neither parent has a bachelor's degree or higher). MU is considered a highly competitive institution, and this is reflected in the academic background of students. The average high school GPA is 3.82 on a 4.0 scale, 72 percent took calculus in high school, and they performed at the 70th percentile on the quantitative section of the SAT or ACT, on average.

In general, the female and male students in the sample are similar in their mean characteristics. The women in the sample are less likely than men to be White (63 vs. 66 percent) and more likely to be Black (3 vs. 2 percent), though the differences are small. Men are slightly more likely to take the course in their first year (77 vs. 76 percent) and third year (4 vs. 3 percent), while women are more likely to take it in their second year (20 vs. 17 percent). Female economics students seem to have lower family incomes (more likely to be in the lowest category and less likely to be in the highest category), but they are also more likely to not have reported income. Notably, women have higher high school GPAs (3.84 vs. 3.81), but men are slightly more likely to have taken calculus (73 vs. 71 percent) and have higher standardized quantitative test scores (74th vs. 65th percentile).

2.5.2 Evidence of Policy Change

I first present descriptive evidence that the stated changes to the grading curve in Principles of Economics I courses did in fact change the distribution of grades instructors awarded. Panel (a) of Figure 2.1 shows the distribution of grades students received by whether they took the course before the curve changed or after, with fall 2016 the first semester under the post regime. From spring 2013 to spring 2016, 31 percent of grades awarded were in the A range (A-, A, or A+); from fall 2016 to spring 2018, 42 percent of grades were some type of A. After the curve changed, instructors gave fewer E, D, or C grades (31 percent pre vs. 19 percent post) and more B and A grades (69 vs. 81 percent).

Panel (b) of Figure 2.1 shows variation in grades awarded at the lecture section level, where a section is a unique instructor, term, and course catalog number. (Most instructors teach one section a semester, and some teach two.) This figure plots the distribution of the proportion of students in a section receiving A grades. Even under the same official curve, there is variation in how many A's instructors award, but the distribution clearly shifts right under the new policy. In the pre-period, the average proportion of A grades was 30 percent,

[6]Because the income categories collected by the university have changed over time, I cannot disaggregate the top category into smaller bins.

but the proportion ranged from 25 to 52. In the post period, the average proportion of A grades was 41 percent, with a range of 32 to 57. Note that both panels of Figure 2.1 suggest that under both policies, instructors were somewhat more generous than what was written in their syllabuses, which guaranteed 25 percent of students some type of A in the pre-period and at least 35 percent in the post period.

2.5.3 Causal Effect of Higher Letter Grades

Table 2.2 presents the main findings, the estimated effect of receiving a higher letter grade in Principles I on subsequent economics course-taking and major declaration. The six outcome variables are indicators for whether a student took Principles of Economics II (the second course in the economics sequence), took Intermediate Microeconomic Theory (the third course in the sequence), or declared an economics, business, STEM, or non-economics social science major, all measured as ever doing so in the two years after they took Principles I. The "treatment" of a higher letter grade is for one higher grade on a scale with pluses and minuses. For example, an A- is one grade higher than a B+.

The first column of Table 2.2 shows effects estimated on the full sample of students who took Principles of Economics I. I find that receiving a higher letter grade in Principles I makes students 2.5 percentage points more likely ($p<0.05$) to take Principles II, the next course in the sequence. The effect on taking Intermediate Micro Theory is a statistically insignificant 0.6 percentage points; the effect on declaring an economics major is similar in magnitude at 0.8 percentage points and marginally statistically significant ($p<0.1$). In terms of non-economics major outcomes, it appears that higher grades in introductory economics make students more likely to major in business, by 2.3 percentage points ($p<0.01$). At MU, taking Principles I and II in the economics department is required for business majors. Paired with the fact that the business major is considered more selective and is associated with higher earnings, it is not surprising that awarding higher economics grades increases the rate of students majoring in business more than the rate majoring in economics. I detect no change in the rate of STEM majoring (a statistically insignificant -0.3 percentage points). The increase in economics and business majors may correspond to a decrease in social science majors by 0.7 percentage points ($p<0.1$). Since the way I identify majors is not mutually exclusive, the remainder of the increase in business and economics could correspond to a decrease in all other majors such as humanities, arts, and communications, or to an increase in the rate of double-majoring. In a separate analysis (not shown) I find no increase in the likelihood of having two or more declared majors, implying students are not adding economics or business as second majors but rather shifting from other fields.

The final three columns of Table 2.2 show effects estimated separately for women and men

and a test for equality between the groups. I find no evidence of heterogeneity by gender in the response to higher grades. The point estimates for women and men are consistently very similar, and I cannot reject the hypothesis that they are equal for any of the outcomes. For example, a higher grade makes women 2.7 percentage points more likely to take Principles II, and men 2.4 percentage points more likely (p-value for difference: 0.806). The effect on declaring a business major is 2 percentage points for women and 2.6 percentage points for men (p-value for difference: 0.517).

2.5.4 Effect of Higher Letter Grade, by Grade

For statistical power reasons, my preferred specification estimates a constant linear effect of receiving a higher letter grade. However, the marginal effect of receiving, say, an A- over a B+ could be different than receiving a C over a C-. This could be true for multiple reasons. Perhaps students psychologically value grades in the A-range, or employers only care about grades above a certain threshold. Both the economics department and the business school require a C grade or higher in Principles I, so the marginal effect of a C might be particularly salient. To investigate this, I estimate a specification similar to Equation 2.1, but with indicators for each grade (A+, A, A-, B+, B, B-, C+, and C, with C- or below the omitted category) rather than a single grade variable. I present estimated effects of each grade relative to the grade just below it, since the "treatment" can be thought of as increasing a student's grade on the margin (and this is the analog of the linear effect). For example, the effect of receiving an A- is relative to a B+, and is calculated as the coefficient on A- minus the coefficient on B+.

Table 2.3 shows the effect of each letter grade (relative to the grade below) on each of the six outcomes, for all students and separately by gender. For economics course-taking, the largest effects of grades on taking Principles II are of receiving an A- (4.6 percentage points, $p<0.05$), a B (4.4 percentage points, $p<0.05$), and a C (5 percentage points, $p<0.05$). I find a marginally significant effect of receiving a C on taking Intermediate Micro (2.8 percentage points, $p<0.1$) and a significant effect of a C on declaring an economics major (2.7 percentage points, $p<0.05$). The effect of a B on majoring in economics is a marginally significant 2.5 percentage points. Though the comparison of magnitudes is consistent with students particularly valuing A range grades and needing a C to meet major requirements, the confidence intervals around the estimates are wide enough that I can't make precise comparisons.

Turning to the effects on declaring a business major, the largest and statistically significant effects appear for relatively lower letter grades (C, B-, and B, all with effects of around 4 percentage points). I find no effect of any grade on majoring in STEM, and the

only significant grade effect on majoring in social science is of a B- (-2.9 percentage points, $p<0.05$).

In terms of gender differences, I find few significantly different effects by gender and no clear pattern. Furthermore, I am conducting many hypothesis tests. The tests for heterogeneity by gender imply that women react more positively to an A+ grade than men in terms of taking Intermediate Micro (8.3 vs. 0.2 percentage points, p-value for difference: 0.095), while men increase their probabilities of taking intermediate micro and declaring an economics major more in response to a B- (5.9 vs. -2.9 percentage points, p-value for difference: 0.006). The effects of an A (4.6 percentage points for women, -2.1 for men, p-value for difference: 0.041) and a B (5.3 for women, 0.5 for men, p-value for difference: 0.095) on declaration of an economics major are more positive for women, but the opposite is true for a B- grade (-1.1 for women, 4.4 for men, p-value for difference: 0.066). Panel B of Table 2.3 suggests that grades of A and A- may induce more men into business relative to women. Consistent with the results in Table 2.2, I interpret all of this as providing no strong evidence that women are particularly responsive to grades.

2.6 Alternative Specifications and Robustness

2.6.1 Measuring Performance as Percentage Points vs. Percentile

Thus far, in controlling for students' raw performance in Principles I, I have used a measure of their percentile rank within their course. The advantage of operationalizing performance as percentile rank is that it has the same meaning and distribution across classes and explicitly maps to grade curve policies, which specify a certain percent of students to receive certain grades. Because instructors vary in the raw scores they give out, students with the same raw score in different courses might fall in very different parts of their course distribution—and therefore the grade distribution. However, one argument against using percentile rank is that students generally do not observe this measure; rather, they observe their total points or percent score. Conceptually, controlling for percent better captures the signal value of the letter grade over and above the information contained in their raw score. In practice, percentile rank and percentage score are monotonic transformations between each other within a course. (This is not necessarily true across courses, but by including instructor fixed effects I account for different grading norms across instructors.)

To test the sensitivity of my results to using percentile rank versus percent score, I present results controlling for the latter in Table 2.4. These results are equivalent to Table 2.2 and estimate Equation 2.1, except with percent score (out of 100) rather than percentile. The estimated effects of higher grades are qualitatively similar across the two specifications,

but conditioning on percent score rather than percentile results in somewhat larger point estimates and smaller standard errors. The first column of results finds that receiving a higher grade in Principles I makes students 4 percentage points more likely to take Principles II, 2.3 percentage points more likely to take Intermediate Micro, and 2.1 percentage points more likely to major in economics, all significant at the $\alpha = 0.01$ level. Higher grades in introductory economics make students 3.2 percentage points more likely to major in business and 1 point less likely to declare a non-economics social science major.

In this specification, I find some evidence of heterogeneity by gender, but in a way that suggests men are more likely to continue in economics as a result of a higher grade. The effect of a higher grade on taking Intermediate Micro is 2.9 percentage points for men and an insignificant 1 percentage point for women (p-value for difference: 0.013). Similarly, the effect on declaring an economics major is 2.5 points for men and an insignificant 0.9 points for women (p-value for difference: 0.018). Tables 2.2 and 2.4 together suggest that, at best, higher grades do not close gender gaps in economics persistence, and at worst they may exacerbate gaps.

2.6.2 Dropping Sections with Grade Rank Inconsistencies

My identification strategy hinges on controlling for students' underlying course performance, which I argue captures the characteristics and inputs that could affect both letter grade and academic choices. This requires a measure of students' final total score in Principles I, which comes from what instructors have entered in learning management system gradebook data. To calculate final score and final percentile, in some cases I could use a "final score" grade entered by instructors with no additional cleaning. In other cases, there was no final score, so I constructed final scores based on individual assignment scores and the weighting of assignments detailed in course syllabuses.

I checked for agreement between the gradebook data and official transcript data by comparing the ranking of raw scores to the ranking of official letter grades. If scores were entered correctly and completely by the instructors and (as needed) calculated correctly by the researcher, then the ranking of raw scores within a class should align with the ranking of letter grades. For example, a student whose final score was an 80 should have a (weakly) lower letter grade than a student with an 85, conditional on instructor, term, and class section. I do find some violations of this rank restriction. Out of 49 sections, 22 have rank inconsistencies between final score and final letter grade.

These inconsistencies are likely due to measurement error; this could occur if for example if an instructor did not input all assignment scores, or calculated final scores outside of the LMS and changed their weighting scheme from what appeared on the initial syllabus.

However, it could also indicate some instructors changing final grades, possibly because of some advocacy on the part of students. The latter case is more problematic for identification because it suggests that students with the same performance could have different grades because of manipulation by the student rather than randomness in the grading curve. The case of measurement error could also bias results because it would mean I am less precisely controlling for underlying performance and effort, which are positively correlated with raw score and grade. I test for robustness to excluding these potentially problematic sections below.

Table 2.5 shows grade effects estimated only on the sample of course sections with rank consistency between raw scores and letter grades. Although I lose power with a smaller sample, the results are generally robust to this sample restriction. I still find a significant effect of higher grade on taking Principles II (3.8 percentage points, $p<0.01$) and on declaring a business major (1.9 percentage points, $p<0.01$). The estimated effect on declaring an economics major is no longer significant, but the point estimate of 0.4 percentage points is similar to the estimate of 0.8 points with the full sample. The small, negative, marginally significant effect of higher grades on social science major declaration also disappears with this sample: the point estimate is a precise 0.00 percentage points and not significant. Like in Table 2.2, I detect no significant differences by gender.

2.6.3 Change to Business School Admissions

One potentially confounding policy change occurred during the same period. At MU, the economics department and the business school are closely related, sharing faculty and students. Undergraduate business majors are required to take introductory micro- and macroeconomics in the economics department, so many students who take Principles are aspiring business majors. Traditionally, students applied for admission to the business major after they had already enrolled at MU. Starting with the entering class of fall 2017, the business school started admitting the majority of its students as pre-admits, meaning they applied as seniors in high school and arrived on campus as already declared business majors. This changed the default major from undeclared to business for a number of students in the sample. In addition, the business school expanded its undergraduate class size. These changes could affect the analysis in several ways. The class size expansion could bias effects on business majoring upwards. On the other hand, the fact that students are already in the business school could potentially make them less responsive to grades, since they are no longer competing to get in (though they must still achieve minimum grades in Principles of Economics).

To see how much this business admissions policy change is affecting the results, I estimate

Equation 2.1 using only observations prior to the 2017-2018 school year. Since the economics grading policy changed in 2016, this still leaves one year of data after the curve change. The results are in Table 2.6. The estimated effects are similar to the effects using the full time period. Notably, when I exclude years affected by the new business school admissions policy, the effect of economics grades on economics majoring is slightly higher (1.1 percentage points compared to 0.8 in Table 2.2) and the effect on business majoring is somewhat lower (1.5 compared to 2.3 percentage points), though the confidence intervals from the two sets of estimates overlap. I conclude that the change to the business school is not responsible for the substantive findings.

2.6.4 Checking for Selection into Grades and Courses

My identification strategy assumes that after controlling for underlying performance, letter grades in Principles I are orthogonal to subsequent academic decisions. If there are factors affecting students' grades which are not captured by the raw measure of performance and which also affect outcomes, this exogeneity assumption would be violated. To test for this type of selection into letter grades, I examine whether, conditional on underlying performance, instructor, academic year, and time of year, letter grade predicts observable characteristics such as gender, race, family background, and academic preparation. This is analogous to testing for discontinuities at the cutoff in observable characteristics in a regression discontinuity setting. Although I control for a rich set of observable characteristics in all of the above analyses, significant relationships between grades and observable characteristics could suggest a relationship between grades and unobservable characteristics, which may be biasing my effects upwards.

I estimate:

$$\begin{aligned} X_{ijt} = \beta_0 + \beta_1 Grade_{ijt} + \beta_2 Percentile_{ijt} + \beta_3 Percentile_{ijt}^2 + \beta_4 Percentile_{ijt}^3 \\ + \sum_j \alpha_j Instructor_j + \delta Fall_t + \lambda Year_t + \epsilon_{ijt} \end{aligned} \tag{2.2}$$

For X variables including gender (female indicator), race (indicators for White, Black, Hispanic, and Asian), high school GPA, whether the student took calculus in high school, score on the university's math placement test, and parent education (whether they have a parent with a graduate or professional degree). All other terms are defined as before.

Table 2.7 shows the results of these falsification tests. Conditional on performance, instructor, and term, students with higher grades are no more likely to be women. A higher grade is associated with a 0.9 percentage point lower chance of a student being Black ($p<0.01$). Higher grades do not predict whether a student took calculus in high school, but

they do predict higher high school GPA; one higher letter grade is associated with 0.008 high school grade points on a 4.0 scale. Higher grades also predict a student's performance on MU's math placement test, by 0.2 points on a 25-point scale. Finally, students with higher grades are somewhat more likely to have a parent with a graduate degree (1.3 percentage points, $p<0.1$).

These results suggest that even conditional on performance, students who receive higher letter grades are different than those with lower grades. The conditional exogeneity assumption may not be fully satisfied and the effects of letter grades I estimate may be upwardly biased. However, the sizes of the associations in Table 2.7 are substantively small, and, once multiplied by correlations between the characteristics and outcomes, unlikely to account for treatment effects of the magnitudes I find.

2.7 Conclusion

Many economics and STEM departments, as they consider ways to attract and retain students and improve representation by gender and other dimensions, are thinking about using grading and evaluation systems as a policy tool that could be used to achieve these goals. For example, the economics department at Duke University instituted a pass/fail grading system for its introductory courses in 2019, motivated by a desire to make the major "more welcoming to students" (Li, 2019). Furthermore, the policy change examined in the current study was partially motivated by prior work suggesting women may be particularly deterred by poor grades.

Using grades as an effective policy lever requires understanding how students' academic decisions change in response to grades and how those responses vary across groups. However, grades are not randomly assigned, so estimating causal effects can be challenging using most observational data. To overcome this challenge, I implement an identification strategy that controls for students' raw, continuous performance in introductory economics courses, which captures much of the often unobservable inputs that determine grades, including effort, motivation, and academic preparation. I exploit variation in letter grades conditional on underlying performance, which comes from both naturally occurring variation in grade cutoffs across semesters as well as a discrete change to the grading curve in introductory economics.

I find that receiving a higher grade in Principles of Economics I - Microeconomics makes students 2.5 percentage points more likely to take the next course in the sequence, Principles of Economics II - Macroeconomics. The effect of eventually declaring an economics major is small at 0.8 percent, and marginally statistically significant. However, I find a substantial

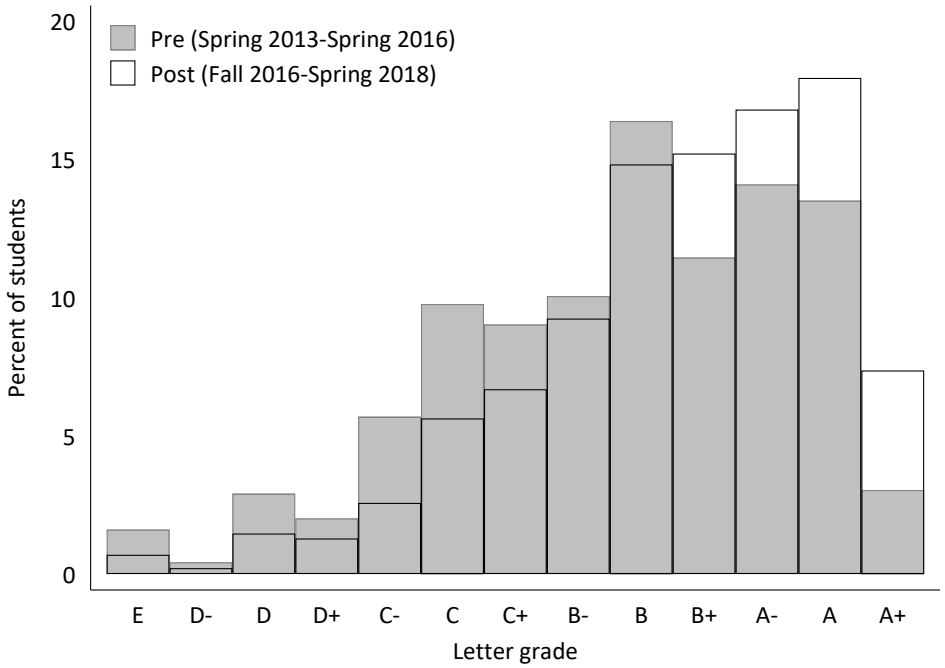
increase—2.3 percentage points—in the probability of declaring a business major, suggesting that higher economics grades allowed students to gain admission to the prestigious business school.

According to Census Bureau data on postsecondary labor market outcomes, business majors at the studied institution have the highest earnings of any major ten years after graduating—more than economics and even more than engineering and computer science majors. While the median MU economics graduate earns $112,000 ten years after graduating, the median business graduate makes over $160,000.[7] This implies that while the economics department may not have succeeded in its goal of attracting more economics majors, giving out higher economics grades likely improved the labor market outcomes of students by giving them access to the highly selective business school.

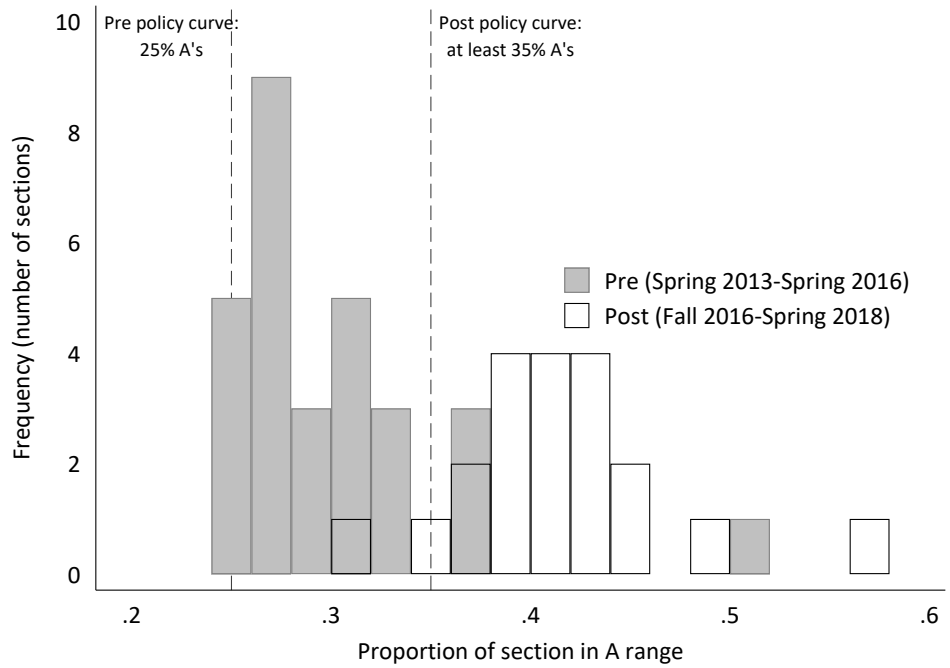
I find very little support for the hypothesis that women respond differently to letter grades than men. In my primary specification, the effect of grades on course-taking and major outcomes are very similar and statistically indistinguishable by gender. In an alternate specification where I control for a student's percent score rather than their percentile ranking, I find differential effects on economics course-taking and major choice, but in the direction of *men* changing their behavior more. The combined results suggest that at best, giving out higher grades in introductory economics did not affect the gender gap in economics; at worst, it could have widened the gap. Economics departments interested in closing gender gaps would be better off pursuing policies that have proved more effective, such as role model interventions (Porter and Serra, 2019) and providing more information about the field of economics (Li, 2018; Bayer et al., 2019).

[7]Based on publicly available Post-Secondary Employment Outcomes (PSEO) data from the U.S. Census Bureau, using the 2004-2006 graduating cohorts.

Figure 2.1: Grade Distributions Pre- and Post-Grading Curve Policy Change



(a) Distribution of Letter Grades (Student Level)



(b) Distribution of Proportion of A Range Grades Awarded (Section Level)

Table 2.1: Sample Descriptive Statistics

	All	Women (W)	Men (M)	p-value, W vs. M
Female	0.39	1.00	0.00	
White	0.65	0.63	0.66	0.00
Asian	0.19	0.20	0.19	0.15
Black	0.03	0.03	0.02	0.00
Hispanic	0.05	0.05	0.05	0.87
Native American, Hawaiian, or Pacific Islander	0.00	0.00	0.00	0.72
Multiple races	0.03	0.03	0.03	0.59
Race/ethnicity missing	0.05	0.05	0.05	0.97
First year at MU	0.77	0.76	0.77	0.05
Second year	0.18	0.20	0.17	0.00
Third year	0.04	0.03	0.04	0.00
Fourth+ year	0.02	0.02	0.02	0.90
Family income less than $25,000	0.03	0.03	0.03	0.02
$25,000-$49,999	0.05	0.05	0.05	0.13
$50,000-$74,999	0.05	0.05	0.05	0.51
$75,000-$99,999	0.07	0.07	0.07	0.92
$100,000 and above	0.53	0.51	0.54	0.00
Family income missing	0.23	0.25	0.22	0.00
Max parent ed less than high school	0.01	0.01	0.01	0.12
High school	0.03	0.04	0.03	0.21
Some college	0.04	0.03	0.04	0.46
Bachelor's degree	0.25	0.25	0.25	0.69
Graduate or professional degree	0.58	0.59	0.58	0.47
Parent education missing	0.09	0.08	0.09	0.13
High school GPA	3.82	3.84	3.81	0.00
HS GPA missing	0.11	0.10	0.12	0.00
Took calculus in high school	0.72	0.71	0.73	0.03
SAT or ACT math percentile	70.41	65.22	73.77	0.00
Missing test score	0.13	0.12	0.14	0.00
N	11,836	4,592	7,244	

Table 2.2: Estimated Effect of Higher Letter Grade in Introductory Economics

Effect of higher grade on:	All	Women (W)	Men (M)	p-value, W vs. M
Took Principles II	.025***	.027***	.024***	0.806
	(0.006)	(0.010)	(0.008)	
Took Interm. Micro	.006	.005	.006	0.839
	(0.005)	(0.007)	(0.007)	
Declared Econ Major	.008*	.009	.007	0.866
	(0.004)	(0.006)	(0.006)	
Declared Business Major	.023***	.02**	.026***	0.517
	(0.005)	(0.008)	(0.006)	
Declared STEM Major	-.003	.002	-.005	0.504
	(0.006)	(0.008)	(0.007)	
Declared Social Science Major	-.007*	-.007	-.007	1.000
	(0.004)	(0.007)	(0.005)	
N	11,836	4,592	7,244	

Notes: $^*p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$. Robust standard errors reported. The results in each column come from a regression of the outcome on a linear term for letter grade. The effects are of one higher letter grade, e.g. going from a B+ to an A- or a C to a C+. Regressions also control for a cubic in percentile rank in the course, gender, race, class standing, family income, parental education, SAT and ACT scores, high school GPA, taking calculus in high school, score on the university's math placement test, instructor, academic year, and season of course (fall vs. spring term). Effects for women and men are estimated in separate regressions. All outcomes are measured in the two years following the term the student took Principles of Economics I. Major categories based on 2-digit CIP codes.

Table 2.3: Estimated Effects of Letter Grades in Introductory Economics, with Separate Effects by Grade

Panel A: Economics Course-taking and Major Outcomes

	Took Principles II				Took Interm. Micro				Declared Econ Major			
	All	Women (W)	Men (M)	p-value, W vs. M	All	Women (W)	Men (M)	p-value, W vs. M	All	Women (W)	Men (M)	p-value, W vs. M
(A+)	.017	.035	.012	0.683	.022	.083**	.002	0.095	-.003	.034	-.015	0.251
	(0.025)	(0.048)	(0.030)		(0.021)	(0.042)	(0.025)		(0.018)	(0.037)	(0.021)	
(A)	-.02	-.049	-.007	0.334	0	.036	-.015	0.171	0	.046*	-.021	0.041
	(0.020)	(0.037)	(0.024)		(0.017)	(0.031)	(0.020)		(0.015)	(0.027)	(0.018)	
(A-)	.046**	.04	.047**	0.851	.02	.052**	0	0.100	.013	.037*	-.002	0.179
	(0.018)	(0.031)	(0.023)		(0.016)	(0.024)	(0.020)		(0.014)	(0.022)	(0.018)	
(B+)	.011	.028	-.002	0.432	-.024	-.034	-.02	0.654	-.015	-.034	-.003	0.282
	(0.018)	(0.030)	(0.023)		(0.016)	(0.024)	(0.021)		(0.014)	(0.022)	(0.019)	
(B)	.044**	.048	.043*	0.898	.014	.042**	-.007	0.114	.025*	.053***	.005	0.095
	(0.020)	(0.031)	(0.026)		(0.016)	(0.021)	(0.023)		(0.015)	(0.020)	(0.021)	
(B-)	.032	.047	.018	0.507	.022	-.029	.059**	0.006	.02	-.011	.044*	0.066
	(0.022)	(0.033)	(0.029)		(0.017)	(0.021)	(0.024)		(0.015)	(0.019)	(0.023)	
(C+)	-.002	-.027	.02	0.314	-.003	-.002	0	0.942	.015	.011	.023	0.670
	(0.023)	(0.034)	(0.032)		(0.017)	(0.022)	(0.025)		(0.015)	(0.019)	(0.022)	
(C)	.05**	.06*	.041	0.690	.028*	.008	.043*	0.268	.027**	.017	.036*	0.468
	(0.024)	(0.034)	(0.034)		(0.016)	(0.020)	(0.025)		(0.013)	(0.016)	(0.021)	

Notes: $^*p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$. Robust standard errors reported. The results in each column come from a regression of the outcome on indicators for each letter grade, with C- or below as the omitted category. The effect of each grade is relative to the grade below it. For example, the effect of an A+ is relative to an A, calculated by subtracting the coefficient on A from the coefficient on A+. Regressions also control for a cubic in percentile rank in the course, gender, race, class standing, family income, parental education, SAT and ACT scores, high school GPA, taking calculus in high school, score on the university's math placement test, instructor, academic year, and season of course (fall vs. spring term). Effects for women (N=4,592) and men (N=7,244) are estimated in separate regressions. Total N=11,836. All outcomes are measured in the two years following the term the student took Principles of Economics I.

Table 2.3 *(continued)*: Estimated Effects of Higher Letter Grades in Introductory Economics, with Separate Effects by Grade

Panel B: Non-Economics Major Choice Outcomes

	Declared Business Major				Declared STEM Major				Declared Social Science Major			
	All	Women (W)	Men (M)	p-value, W vs. M	All	Women (W)	Men (M)	p-value, W vs. M	All	Women (W)	Men (M)	p-value, W vs. M
(A+)	-.019	-.073	-.001	0.192	-.026	.022	-.049*	0.191	-.011	-.009	-.012	0.855
	(0.024)	(0.048)	(0.027)		(0.025)	(0.046)	(0.029)		(0.008)	(0.016)	(0.009)	
(A)	-.001	-.053	.018	0.089	-.014	.003	-.021	0.552	0	-.003	.002	0.775
	(0.019)	(0.036)	(0.022)		(0.019)	(0.033)	(0.022)		(0.007)	(0.015)	(0.008)	
(A-)	.026	-.041	.06***	0.004	-.021	.006	-.031	0.293	.005	.017	-.001	0.316
	(0.017)	(0.030)	(0.020)		(0.017)	(0.028)	(0.021)		(0.008)	(0.015)	(0.009)	
(B+)	.029*	.029	.03	0.986	-.001	.014	-.006	0.548	-.007	.005	-.016	0.276
	(0.015)	(0.027)	(0.018)		(0.016)	(0.026)	(0.021)		(0.009)	(0.017)	(0.010)	
(B)	.041***	.038	.044**	0.846	.006	-.021	.024	0.182	-.008	.005	-.017	0.352
	(0.015)	(0.026)	(0.018)		(0.017)	(0.025)	(0.023)		(0.011)	(0.020)	(0.014)	
(B-)	.041***	.059**	.026	0.241	-.015	-.019	-.009	0.810	-.029**	-.022	-.036*	0.638
	(0.014)	(0.023)	(0.017)		(0.019)	(0.027)	(0.026)		(0.015)	(0.023)	(0.019)	
(C+)	.008	.002	.013	0.696	.005	.008	.005	0.940	.001	-.018	.016	0.319
	(0.014)	(0.022)	(0.018)		(0.020)	(0.027)	(0.029)		(0.017)	(0.025)	(0.022)	
(C)	.045***	.067***	.035**	0.223	-.008	-.009	-.008	0.989	-.009	-.019	-.001	0.628
	(0.013)	(0.021)	(0.017)		(0.021)	(0.029)	(0.030)		(0.019)	(0.029)	(0.024)	

Notes: $^*p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$. Robust standard errors reported. The results in each column come from a regression of the outcome on indicators for each letter grade, with C- or below as the omitted category. The effect of each grade is relative to the grade below it. For example, the effect of an A+ is relative to an A, calculated by subtracting the coefficient on A from the coefficient on A+. Regressions also control for a cubic in percentile rank in the course, gender, race, class standing, family income, parental education, SAT and ACT scores, high school GPA, taking calculus in high school, score on the university's math placement test, instructor, academic year, and season of course (fall vs. spring term). Effects for women (N=4,592) and men (N=7,244) are estimated in separate regressions. Total N=11,836. All outcomes are measured in the two years following the term the student took Principles of Economics I. Major categories based on 2-digit CIP codes.

Table 2.4: Estimated Effect of Higher Letter Grade in Introductory Economics, Controlling for Percent Score Rather than Percentile

Effect of higher grade on:	All	Women (W)	Men (M)	p-value, W vs. M
Took Principles II	.04***	.038***	.039***	0.882
	(0.005)	(0.008)	(0.006)	
Took Interm. Micro	.023***	.01	.029***	0.013
	(0.004)	(0.006)	(0.005)	
Declared Econ Major	.021***	.009	.025***	0.018
	(0.003)	(0.005)	(0.005)	
Declared Business Major	.032***	.041***	.026***	0.059
	(0.004)	(0.006)	(0.004)	
Declared STEM Major	.003	.007	.003	0.657
	(0.004)	(0.007)	(0.006)	
Declared Social Science Major	-.01***	-.008	-.011***	0.686
	(0.003)	(0.005)	(0.003)	
N	11,836	4,592	7,244	

Notes: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$. Robust standard errors reported. The results in each column come from a regression of the outcome on a linear term for letter grade. The effects are of one higher letter grade, e.g. going from a B+ to an A- or a C to a C+. Regressions also control for a cubic in percent score in the course, gender, race, class standing, family income, parental education, SAT and ACT scores, high school GPA, taking calculus in high school, score on the university's math placement test, instructor, academic year, and season of course (fall vs. spring term). Effects for women and men are estimated in separate regressions. All outcomes are measured in the two years following the term the student took Principles of Economics I. Major categories based on 2-digit CIP codes.

Table 2.5: Estimated Effect of Higher Letter Grade in Introductory Economics, Dropping Course Sections with Grade Rank Inconsistencies

Effect of higher grade on:	All	Women (W)	Men (M)	p-value, W vs. M
Took Principles II	.038***	.043***	.032***	0.563
	(0.009)	(0.014)	(0.012)	
Took Interm. Micro	.004	-.005	.01	0.295
	(0.007)	(0.010)	(0.011)	
Declared Econ Major	.004	.006	.002	0.773
	(0.007)	(0.009)	(0.010)	
Declared Business Major	.019***	.018	.023**	0.726
	(0.007)	(0.011)	(0.009)	
Declared STEM Major	-.004	.001	-.007	0.598
	(0.008)	(0.012)	(0.012)	
Declared Social Science Major	0	0	0	0.999
	(0.006)	(0.011)	(0.008)	
N	6,902	2,829	4,073	

Notes: *$p < 0.1$, **$p < 0.05$, ***$p < 0.01$. Robust standard errors reported. Sample excludes course sections for which the ranking of raw score (from learning management system data) is inconsistent with the ranking of final letter grade (from university transcript data). The results in each column come from a regression of the outcome on a linear term for letter grade. The effects are of one higher letter grade, e.g. going from a B+ to an A- or a C to a C+. Regressions also control for a cubic in percentile rank in the course, gender, race, class standing, family income, parental education, SAT and ACT scores, high school GPA, taking calculus in high school, score on the university's math placement test, instructor, academic year, and season of course (fall vs. spring term). Effects for women and men are estimated in separate regressions. All outcomes are measured in the two years following the term the student took Principles of Economics I. Major categories based on 2-digit CIP codes.

Table 2.6: Estimated Effect of Higher Letter Grade in Introductory Economics, Excluding 2017-18 Observations

Effect of higher grade on:	All	Women (W)	Men (M)	p-value, W vs. M
Took Principles II	.018***	.021*	.015*	0.666
	(0.007)	(0.011)	(0.009)	
Took Interm. Micro	.008	.004	.011	0.558
	(0.005)	(0.007)	(0.007)	
Declared Econ Major	.011**	.008	.013**	0.547
	(0.005)	(0.007)	(0.007)	
Declared Business Major	.015***	.014	.017**	0.776
	(0.005)	(0.009)	(0.007)	
Declared STEM Major	-.002	.004	-.004	0.527
	(0.006)	(0.009)	(0.008)	
Declared Social Science Major	-.005	-.006	-.004	0.889
	(0.004)	(0.008)	(0.005)	
N	9,102	3,486	5,616	

Notes: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$. Robust standard errors reported. The results in each column come from a regression of the outcome on a linear term for letter grade. The effects are of one higher letter grade, e.g. going from a B+ to an A- or a C to a C+. Regressions also control for a cubic in percentile rank in the course, gender, race, class standing, family income, parental education, SAT and ACT scores, high school GPA, taking calculus in high school, score on the university's math placement test, instructor, academic year, and season of course (fall vs. spring term). Effects for women and men are estimated in separate regressions. All outcomes are measured in the two years following the term the student took Principles of Economics I. Major categories based on 2-digit CIP codes. Sample includes only students who took Principles I prior to the 2017-18 school year.

Table 2.7: Falsification Test: Does Letter Grade Predict Student Characteristics, Conditional on Performance, Instructor, Year, and Season

	Coefficient on grade in Principles I	N in regression
Female	-0.000	11,836
	(0.007)	
White	0.008	11,203
	(0.007)	
Asian	-0.001	11,203
	(0.006)	
Black	-0.009***	11,203
	(0.003)	
Hispanic	0.000	11,203
	(0.003)	
High School GPA	0.008***	10,529
	(0.003)	
Took calculus in high school	0.007	11,836
	(0.006)	
Math placement score (out of 25)	0.213***	11,357
	(0.064)	
Parent has graduate degree	0.013*	10,814
	(0.007)	

Notes: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$. Robust standard errors reported. Each coefficient is from a regression of the student characteristic on a linear term for letter grade, controlling for a cubic in percentile rank in the course, instructor, academic year, and season of course (fall vs. spring term). Observations missing a value for the characteristic are not included.

CHAPTER III

The Advanced Placement Program and Educational Inequality

3.1 Introduction

Since its introduction in 1952, the Advanced Placement (AP) program, which provides an opportunity for ambitious high school students to take college-level courses and possibly obtain college credit or placement out of introductory college courses, has grown dramatically and become nearly ubiquitous in American high schools. Despite the popularity of AP and the perception that participation improves college preparation, increases chances of college admission, and accelerates degree attainment, there is little convincing causal evidence on how taking AP courses affects human capital investment and later outcomes.

In this paper, I investigate whether and how Advanced Placement courses affect the selectivity of college students attend, on-time college graduation, and overall college graduation. I use administrative data from the state of Michigan and exploit variation within high schools across time in how many AP courses are offered to identify the causal effect of AP course availability, AP course-taking, and AP exam-taking. Because there is reason to believe that AP affects different types of students differently, I also test for heterogeneous treatment effects by socioeconomic status, race, and prior academic preparation level. My results suggest that higher income students, White and Asian students, and higher-achieving students are both more likely to take advantage of AP courses when they are offered as well as more likely to reap the benefits of taking them.

I find precisely estimated null reduced form effects of AP course availability on college selectivity and degree attainment. However, these average effects mask important heterogeneity. While the availability of an additional AP course increases competitive college enrollment for higher-income students by 0.4 percentage points (1 percent) and four-year degree attainment by 0.3 percentage points (1 percent), the effects for low-income

students are negative; for students eligible for subsidized meals, an additional available AP course at their school decreases selective college enrollment by 0.4 percentage points (2 percent), four-year college graduation by 0.6 percentage points (13 percent), and six-year degree attainment by 0.7 percentage points (5 percent). Similarly, while White and Asian students increase their college outcomes as a result of additional AP course availability, Black, Hispanic, American Indian, and other Native students see lower college graduation rates. Combined with null first-stage effects of AP course availability on AP course-taking for low-income and underrepresented minority students, these negative reduced form effects imply negative spillover effects of expanding AP access, which could occur as a result of the changing composition of peers in an increasingly tracked environment, or from the diversion of resources away from non-AP students, teachers, and classes. The first stage effects are crucial for understanding the reduced form results, but are not sufficiently strong to be used as valid instruments. For this reason, I do not ultimately report any instrumental variables estimates.

The estimated effects are substantively small, changing the outcomes of only a handful of students. Taken as a whole, the results suggest that at best, expanding AP programs will do little to improve outcomes or close achievement gaps; at worst, increasing AP course offerings may slightly harm already disadvantaged students and widen existing inequalities.

Because this strategy relies on the conditional exogeneity of AP course offerings within a school, I test for possible endogeneity of AP course offerings in two ways. First, I split my panel of schools into two shorter panels. Within these shorter time spans, systematic changes to course offerings and student population are less likely, and fluctuations in AP course offerings are more likely due to idiosyncratic shocks. Second, I test whether higher-achieving students (measured by scores on a standardized middle school math test) select into high schools with more AP courses available. The results from both exercises imply that these potential issues are not driving my findings.

The paper proceeds as follows: Section 3.2 provides history and background on the AP program; Section 3.3 lays out a theoretical framework and reviews prior related work; Section 3.4 describes the methodological approach and data; Section 3.5 presents findings about the effect of AP courses on college outcomes, as well as on the first stage of AP course-taking; Section 3.6 discusses threats to identification and presents robustness checks; and Section 3.7 concludes.

3.2 Background

The Advanced Placement (AP) program traces its origins to just after World War II, when the Ford Foundation created the Fund for the Advancement of Education and concluded that better coordination between secondary and postsecondary schools would help increase the number of college entrants and graduates in the United States. A committee was formed "to develop high school course descriptions and assessments that colleges would find rigorous enough to use as a basis for granting credit" and a pilot program in 11 subject areas was launched in 1952 (College Board, 2003). Since 1955, the AP program has been run by the College Board, the same non-profit organization responsible for the SAT college entrance exam.

Participation has grown dramatically since the program's inception, from 1,229 students at 104 schools nationwide in the 1955-56 academic year (the first year data are available from the College Board) to 2.8 million students and nearly 23,000 schools in 2019 (College Board, 2019). In 2012 (the most recent year for which this figure is available), 74 percent of all public high schools offered AP courses (Malkus, 2016), and these schools serve even more students as a proportion of all public high school students (Theokas and Saaris, 2013). Trends for the state of Michigan are similar. Furthermore, a non-trivial amount of federal, state, and local public funds are dedicated to subsidizing AP teacher training, exam fees, and performance incentives (Klopfenstein, 2010). The U.S. Department of Education created Advanced Placement Incentive Program grants in the late 1990s to increase AP participation among low-income students and reduce achievement gaps; this program was expanded under No Child Left Behind in 2001 (Klopfenstein, 2010). In 2016, the Department of Education awarded over $28 million to subsidize exam fees for low-income students in 41 states (including $560,000 to Michigan) plus the District of Columbia (U.S. Department of Education, 2016).

The AP program serves several ostensible purposes. The College Board describes it as a way to "[enable] willing and academically prepared students to pursue college-level studies—with the opportunity to earn college credit, advanced placement or both—while still in high school" (Rodriguez et al., 2013). The College Board also touts participation as beneficial to college admission and performance, saying that "Taking AP courses demonstrates to college admission officers that students have sought the most rigorous curriculum available to them, and research indicates that students who score a 3 or higher on an AP Exam typically experience greater academic success in college and are more likely to earn a college degree than non-AP students" (Rodriguez et al., 2013). As summarized by Klopfenstein and Thomas (2010), "while the College Board generally makes no explicit

statements that AP experience is a cause of college success, their promotional literature readily leads readers to such a conclusion."

As of 2021, the College Board offers 34 AP courses in six subject areas, including science, math and computer science, history and social sciences, English, world languages and cultures, and arts. In 2019, the most popular subjects (by exams taken) were English Language and Composition, U.S. History, English Literature and Composition, U.S. Government, and World History (College Board, 2020).

3.3 Theoretical Background and Related Literature

My research questions fit within a number of overlapping literatures. One way to conceptualize the AP program is as a high-ability track within a high school. There is a large literature on ability and achievement tracking that informs theory about the effects of AP participation, particularly differential effects by student type (see Betts, 2011, for a review). Theoretically, there is an efficiency-equity tradeoff in any tracking system; the empirical evidence is mixed. Others see AP as a type of or alternative to dual enrollment programs (see, for example, Klopfenstein and Lively, 2012), which have the explicit goal of reducing the financial and time cost of postsecondary education. There is also a broader literature on high school curriculum (e.g., Altonji, 1995).

There are several key mechanisms by which we might expect participation in AP courses and exams to affect human capital investment and educational outcomes. At a high level, participation in AP can change the expected costs and benefits of investment in higher education, and therefore students' optimal level of investment (this follows from the canonical Becker-Rosen model of educational investment). Note that educational investment is multi-dimensional and includes choice of college, number of years in school, and choice of major. This framework is borrowed from Jackson (2010), who writes that, "It is useful to think of taking AP courses as a way to reduce college costs (increased likelihood of admission, more financial aid, tuition savings due to college credit, faster graduation, and signal to colleges about ability or motivation)." In this section, I lay out these channels—as well as a few others—along with the behaviors and outcomes we would expect them to change, citing prior research when appropriate.

3.3.1 College Credit and Placement

As the name suggests, one of the primary channels by which taking AP can benefit a student is by earning college credit and/or placement out of introductory-level college courses. The specific policies on this are determined by individual institutions and vary

widely across and even within institutions. In general, a minimum score on an AP exam (three in most cases, though some schools only accept fours or fives) is required for credit or placement. The most obvious outcome this could affect is time to graduation, with students who earn AP credits more likely to graduate in four or even three years. Using a regression discontinuity design that exploits cutoffs in continuous AP exam scores that translate into the 1-5 integer scores reported to students and colleges, Smith et al. (2017) find that receiving a credit-granting score (a 3 in most cases) positively affects on-time college graduation. The credit and placement mechanism could also affect intermediate outcomes such as choice of college, courses, and major, if certain schools grant credit or placement and others dont, or if only particular AP subjects are associated with the benefits. Gurantz (forthcoming) uses a similar RD strategy to examine college course-taking by subject, finding that women who earn credit from AP exams in STEM subjects take more STEM courses.

3.3.2 College Readiness and Achievement

AP courses are generally considered more rigorous than standard high school classes, so that the experience of taking an AP course and preparing for the exam may directly increase students' knowledge, skills, and college readiness—both in a specific subject or subject area and in general. If this is true, we would expect AP participation to raise students' college GPA, persistence, and likelihood of graduating. Although there is a dearth of credible causal research on this mechanism, the College Board emphasizes it, as mentioned in the previous section. Jackson (2010) evaluates the Advanced Placement Incentive Program in Texas, which paid students and teachers for passing AP exams. He exploits exogenous variation in when schools implemented the program and finds that it increased participation in AP courses and the number of students scoring highly on the SAT or ACT. In the only experimental work to date, Conger et al. (2021) randomly assigned high school students into a treatment that included the option to enroll in newly introduced AP Biology or Chemistry course in their schools. Taking an AP science course resulted in a higher self-reported level of course rigor and a higher level of science skill.

The theoretical effect of AP on students' educational outcomes is not unambiguously positive. Whether prior academic preparation and the rigor of AP courses are substitutes or complements will determine who benefits and, in particular, the likely effect of expanding AP access from its current rates (since schools with no or a low number of AP offerings tend to serve more disadvantaged students). Conger et al. (2021) raise the possibility that students with less preparation might be "unable to engage in the material" or get more discouraged; on the other hand, they might have more to gain. Their experimental evidence suggests that students with stronger science preparation increased their science knowledge more as a

result of randomized access to AP science courses.

3.3.3 College Admissions

A third channel, which is closely linked to the above, is that AP participation can serve as a signal of student ability, motivation, and college readiness, as well as a signal of school quality, which are used by college admissions committees in evaluating applicants. Some universities have deterministic formulas for doing this, such as awarding additional points to an application or recalculating GPA upwards for each AP course taken; others have more holistic approaches but take AP into consideration in their evaluation process. If students and parents are aware of this (and anecdotal evidence as well as the College Board's own promotional materials suggests that the benefit of AP for college admissions is highly salient), AP participation could affect the portfolio of colleges students apply to. It would also affect admission conditional on application, and could ultimately affect where students matriculate.[1]

3.3.4 Ability Signaling and Belief Updating

Performance in AP courses and particularly on AP exams may serve as a signal to students that causes them to re-asses their own academic ability and potential for college success (for evidence that grades and standardized test scores can lead to this type of belief updating, see Gonzalez, 2017; Goodman, 2016; and Jacob and Wilder, 2010). Depending on whether they were over- or under-estimating their ability to begin with, students might revise their beliefs upwards or downwards. This belief updating could translate into a number of outcomes, including choice of colleges applied to, matriculation, and subsequent course-taking behavior.

Again, we might expect quite different effects by student characteristics such as income and prior preparation level. Conger et al. (2021) find that more prepared students gain confidence from AP, while less prepared ones lose confidence in their ability to succeed in college-level courses. In a longer-term follow-up to the RCT in Conger et al. (2021), Conger et al. (2020) find a *negative* effect of AP access on college selectivity. The effect appeared for 12th grade students only (not 11th graders), who had already applied to college, suggesting it was not driven by application or admissions, but rather a negative signal that affected matriculation decisions. The negative treatment effect is also concentrated among students who were initially less academically prepared. These somewhat discouraging findings point

[1] I am only considering partial equilibrium effects here. If every student took one additional AP course, the advantage in admissions (given a fixed number of slots) would disappear.

to the importance of considering who the marginal students are when expanding access. In a world where nearly all schools have AP courses available, the marginal student may be less prepared and be made worse off.

This belief-updating mechanism is likely to operate through specific subjects or fields as well as in regards to overall academic ability. Avery et al. (2018) use the same regression discontinuity design as Smith et al. (2017) and find that receiving a higher score on an AP exam significantly increases the likelihood that a student will major in that subject in college; they argue that "a substantial portion of the overall effect is driven by behavioral responses to the positive signal of receiving a higher score."

3.3.5 Other Mechanisms

Related to but distinct from student skill and beliefs about ability is students' interest in academic subjects. AP curricula tend to emphasize critical thinking and inquiry over rote memorization, which Conger et al. (2021) hypothesize might "spur greater interest" in the subject (biology or chemistry in the case of their experiment) "because it becomes more enjoyable and more accessible." Additionally, some AP courses, such as economics, art history, and psychology, are in topics not typically taught in high school; exposure to these fields might spark an interest in a previously unexplored subject and affect students' initial choice of college major. Conger et al. (2021) find that the availability of AP science increases students' interest in pursuing a STEM degree.

As emphasized by Jackson (2010) and Conger et al. (2021), the effort required to succeed in an AP course and pass an AP exam could crowd out effort in other academic and non-academic tasks, depending on the degree of complementarity between the various tasks. This is particularly important in considering policies that subsidize or incentivize AP participation in some way, as they may induce some students to take more than the socially optimal number of AP courses or exams. The Conger et al. experiment found that taking AP science lowered grades in both science and non-science courses, though given previously discussed mechanisms such as skill acquisition and admissions advantages, the net effect could be positive or negative.

The rigor of AP courses might also be a source of stress for students. Conger et al. (2021) cite psychological research documenting a U-shaped relationship between stress and performance and suggest that the pressure of AP may hinder the ability to learn. In their experiment, students experienced an increase in stress levels.

Finally, in a world of limited educational resources, a strong AP program in a school could divert resources away from non-AP courses, decreasing the quality of instruction in other courses. This could negatively affect the learning of both AP students in their other

classes as well as non-AP students. On the other hand, AP teachers and students might have positive spillovers on non-AP students and classes if the training or skills they learn in AP translate to other contexts.

3.4 Method and Data

3.4.1 Identification

Simply comparing students or schools with different levels of AP courses or exams will give an upwardly biased estimate of the effect on educational outcomes, since students taking AP and schools offering AP tend to be higher-achieving to begin with. To estimate the causal impact of AP course availability on college outcomes, I exploit time variation in how many AP courses a high school offered each year. My strategy is similar to that of Darolia et al. (2020), who use what they argue is "plausibly exogenous variation in course offerings within high schools over time" to study the effect of STEM course availability on postsecondary STEM enrollment and degree attainment in Missouri. My identification strategy, like theirs, hinges on year-to-year differences in course offerings within a school being (conditionally) exogenous. This would be the case if the variation is due to things like unrelated changes in teaching staff (due to, e.g., retirement or parental leave) and rules governing class size.

I use panel data covering the graduating classes of 2005 through 2012 in a sample of Michigan public high schools. By controlling for school fixed effects, I compare a cohort of high school seniors to another cohort from the same school, where one cohort had a higher number of AP courses available to them. I also include year fixed effects to account for the general upward trend in both AP and college outcomes. School-specific time trends account for the possibility that schools on an especially steep trajectory in terms of outcomes differentially select into offering more APs.

My primary estimating equation is

$$Y_{ijt} = \beta_0 + \beta_1 AP_{jt,t-1} + \sum_j \delta_j + \sum_{t=2003}^{2012} \lambda_t + \sum_j \tau_j \cdot t + \boldsymbol{\gamma X_i'} + \boldsymbol{\mu Z_{j,t-2}'} + \varepsilon_{ijt} \tag{3.1}$$

where Y_{ijt} is the outcome of interest for student i graduating from school j in year t. The three outcomes I measure are (1) whether a student enrolled at a college that is classified as competitive or higher by the Barron's selectivity index, (2) whether they earned a bachelor's degree within four years of graduating high school, and (3) whether they earned a bachelor's degree within six years. The count variable $AP_{jt,t-1}$ is the number of AP subjects available

to cohort t at school j during their junior and senior year;[2] δ_j are school fixed effects; λ_t are year fixed effects; and τ_j are school-specific linear time trends. The vector $\boldsymbol{X_i}$ includes student characteristics (sex, race, free or reduced-price lunch eligibility in 12th grade, and standardized score on a middle school math test); $\boldsymbol{Z_{j,t-2}}$ captures time-varying school characteristics (average middle school math test score, school size, student-to-teacher ratio, per-student spending, and local unemployment), measured in the student's sophomore year so that they are unaffected by the treatment. I estimate Equation 3.1 with a linear probability model and cluster standard errors at the school level.

To test for heterogeneity by socioeconomic status, I add an interaction term between the number of AP courses and an indicator for eligibility for free or reduced-price lunch (FRPL) in 12th grade.[3] To test for heterogeneity by race, I add an interaction term between the number of AP courses and underrepresented minority (URM) status. To test for heterogeneity by academic preparation, I add an interaction with standardized score on the Michigan standardized math test in middle school.[4] The equations I estimate are:

$$\begin{aligned} Y_{ijt} = {} & \beta_0 + \beta_1 AP_{jt,t-1} + \beta_2 FRPL_{it} + \beta_3 AP_{jt,t-1} \cdot FRPL_{it} \\ & + \sum_j \delta_j + \sum_{t=2003}^{2012} \lambda_t + \sum_j \tau_j \cdot t + \boldsymbol{\gamma X_i'} + \boldsymbol{\mu Z_{j,t-2}'} + \varepsilon_{ijt} \end{aligned} \tag{3.2}$$

$$\begin{aligned} Y_{ijt} = {} & \alpha_0 + \alpha_1 AP_{jt,t-1} + \alpha_2 URM_i + \alpha_3 AP_{jt,t-1} \cdot URM_i \\ & + \sum_j \delta_j + \sum_{t=2003}^{2012} \lambda_t + \sum_j \tau_j \cdot t + \boldsymbol{\gamma X_i'} + \boldsymbol{\mu Z_{j,t-2}'} + \varepsilon_{ijt} \end{aligned} \tag{3.3}$$

$$\begin{aligned} Y_{ijt} = {} & \eta_0 + \eta_1 AP_{jt,t-1} + \eta_2 Math_i + \eta_3 AP_{jt,t-1} \cdot Math_i \\ & + \sum_j \delta_j + \sum_{t=2003}^{2012} \lambda_t + \sum_j \tau_j \cdot t + \boldsymbol{\gamma X_i'} + \boldsymbol{\mu Z_{j,t-2}'} + \varepsilon_{ijt} \end{aligned} \tag{3.4}$$

[2] As an example, if school j offered AP Biology and U.S. History in 2006 and Biology and U.S. Government in 2007, $AP_{j,2007,2006}$ would equal 3. This variable can take values between 0 and 26 AP subjects. I collapsed a number of subjects that the transcript data didn't allow me to distinguish between. For example, microeconomics and macroeconomics are two distinct subjects, but many schools just listed "AP economics." Appendix Figure B.1 summarizes these decisions.

[3] In Michigan, the threshold for subsidized lunch is family income up to 185 percent of the federal poverty line. In 2019, this was equivalent to $47,638 for a family of four.

[4] The grades in which the state of Michigan tests students by subject have changed over time. I use a student's eighth grade test score if it is available, and their seventh grade score if not. I use math scores because the other subject tests were not offered in the relevant years for the full sample.

Here, β_1 measures the effect of an additional course for a higher-income student; the sum of $\beta_1 + \beta_3$ gives the effect for a lower-income student. Similarly, α_1 is the effect for White and Asian students, while $\alpha_1 + \alpha_3$ is the effect for Black, Hispanic, and Native students. η_1 is the effect for a student with an average middle school math score, and $\eta_1 + \eta_3$ is the effect for a student with a math score one standard deviation above the mean.[5]

Equations 3.1 through 3.4 represent the reduced form or intent-to-treat effect of AP course availability. While this is a policy-relevant parameter for schools and districts considering introducing or expanding an AP program, we may also want to know what the effect of an additional AP course is for the students who actually take the course. To do this, I estimate an instrumental variables specification, where the first stage predicts AP course- or exam-taking as a function of AP course availability, and the second stage estimates the effect of AP course- or exam- taking on the outcome, using course availability as an instrument:

$$\begin{aligned}(\#\text{ AP courses taken })_{ijt,t-1} = \alpha_0 + \alpha_1(\#\text{ AP courses available})_{jt,t-1} + \sum_j \delta_j \\ + \sum_{t=2003}^{2012} \lambda_t + \sum_j \tau_j \cdot t + \boldsymbol{\gamma X_i'} + \boldsymbol{\mu Z_{j,t-2}'} + \varepsilon_{jt}\end{aligned} \tag{3.5}$$

$$\begin{aligned}Y_{ijt} = \beta_0 + \beta_1(\#\text{ AP } \widehat{\text{courses taken}})_{ijt,t-1} + \sum_j \delta_j \\ + \sum_{t=2003}^{2012} \lambda_t + \sum_j \tau_j \cdot t + \boldsymbol{\gamma X_i'} + \boldsymbol{\mu Z_{j,t-2}'} + \varepsilon_{jt}\end{aligned} \tag{3.6}$$

In this set-up, the parameter β_1 is the local average treatment effect: the effect of taking an additional AP course or exam, for the student who takes the additional course when it becomes available. To the extent that these marginal students are supply-constrained, we would expect an additional course to have a positive impact. However, if the marginal student would not otherwise take AP but is pushed into an AP course by misinformation or by explicit or implicit incentives, taking the additional course might be sub-optimal and have a negative effect.

To test for treatment effect heterogeneity with the instrumental variables specification, I interact both the instrument (course availability) and the treatment (course- or exam-taking) with the subgroup variable of interest (FRPL status, URM status, or middle school

[5]Test scores are not available for all students; they would be missing if the student attended middle school in a different state or at a private school, or if they were exempt from the test. For the heterogeneity analysis by test score, students missing test scores are omitted. In all other analyses, test score is coded as zero if missing and I include an indicator for missing test score.

test score).

The validity of the IV estimates relies on the exclusion restriction that the presence of AP courses at a school affects students only so far as it encourages them to take more AP courses and exams. This would be violated if some of the spillover-type mechanisms from Section 3.3.5 are at play, such as positive spillovers of AP content and a more college-oriented culture, or negative spillovers due to diversion of resources. The direction of the bias here is theoretically ambiguous. For this reason, I consider the reduced form effects more internally valid. Even if the IV approach is not valid, the first stage—how AP course- and exam-taking changes when courses become available—provides important evidence about how increasing AP offerings increases access, and for whom.

3.4.2 Data

The data I use are provided by the Michigan Consortium for Education Research (MCER) and accessed through the University of Michigan's Education Policy Initiative. My first data source is the MCER Transcript Study, which collected longitudinal transcript data from a random sample of Michigan public high schools. This dataset includes, for each school in the sample, every course taken by students at that school in a given year. The collected course data are sufficiently clean for 87 of the schools.

In order to measure the treatment I am interested in—AP courses available by school and AP courses taken by student—I systematically identified which courses were AP based on course title. The way in which schools list courses is not standardized across schools. Flagging courses as AP was an iterative process that started with more obvious course titles (e.g. "AP Calculus" or "Advanced Placement Biology") and continued by searching for other phrases associated with AP and with one of the recognized AP subjects (e.g. "AP CMP GOV" for comparative government and politics). While some courses were obviously AP, others were more ambiguous. If I wasn't reasonably sure a course was AP, I erred on the more conservative side and did not classify it as AP. I assign course availability at the school level and course-taking at the student level, counting by number of subjects. For a subset of the students for whom I have course-taking data, I can also observe how many AP exams they took. MCER has access to all AP exams taken by Michigan students between 2006 and 2013. Since most students take AP in their junior and senior years, I can count AP exams for the classes of 2007 onward.[6]

To identify cohorts of high school seniors by school, I use the Single Record Student Database and the Michigan Student Data System. This student-by-year panel dataset

[6] As is standard in the education literature, years refer to the spring of the academic year. For example, 2006 refers to the 2005-2006 school year.

contains demographic information (including free and reduced lunch eligibility) as well as the school and district each student attends each year. After keeping 12th grade observations for students in the schools and years covered by the transcript sample, I merge these data with National Student Clearinghouse (NSC) data in order to measure college outcomes. The NSC provides information on college enrollment at any four- or two-year school in the country (with a few exceptions), by date of enrollment and institution.

My final sample includes 174,469 students at who were seniors at 87 public Michigan high schools between 2005 and 2012.

3.5 Results

3.5.1 Descriptive Results

I begin with descriptive statistics about the students and schools in the sample, summarized in Table 3.1. Roughly half of the students are female. The majority, 78 percent, are White, four percent are Asian, 17 percent are Black, three percent are Hispanic, and fewer than one percent are Native (a category which includes American Indian, Alaska Native, Native Hawaiian, and Pacific Islander students). Given the small number of Asian, Hispanic, and Native students, for analyses by race and ethnicity I collapse the categories into underrepresented minority students (Black, Hispanic, and Native) and non-URM (White and Asian). Around a quarter of students in the sample are eligible for free or reduced-price lunch, which I use as a proxy for family income. These means all closely resemble the full population of Michigan seniors during this time. At the school-cohort level, the average school in the panel enrolls around 1400 students, has a student-to-teacher ratio of 21, spends $6,300 per student, and has a local unemployment rate of 9 percent.[7]

The average student in the sample has just under 10 AP courses available to them during their junior and senior year, takes 0.78 courses, and takes 0.73 exams. The average school offers 8.56 courses to a cohort. I provide more detail on the variation in AP course offerings by school and across time, as well as AP course- and exam-taking, in Appendix Table B.1 and Appendix Figures B.2 through B.10. Over time, the most common AP course offerings are English, Calculus, U.S. History, Biology, and Chemistry (see Appendix Table B.1). The most common courses taken are English, Calculus, U.S. Government, Biology, and Psychology; and the most popular exams are English, Calculus, U.S. History, U.S. Government, and Biology.[8]

[7]The minimum values for the school enrollment and student-to-teacher ratio variables are zero. These characteristics are measured two years prior to the current year, so zeros reflect the small number of schools that were recently established.

[8]Recall that English and Calculus are each actually two separate courses: English Literature and

While the vast majority of schools offered at least one AP course to their juniors and seniors over the entire period, there is considerable variation in the number offered. The number of AP courses varies both across and within schools over time (see Appendix Figures B.3, B.4, and B.5), and the changes go in both directions. My identifying variation comes from within-school increases and decreases in AP course offerings. These changes are driven by particular courses. The most common subjects to be introduced are World History, Economics, Psychology, Biology, and Statistics; the most likely to be taken away are Psychology, Computer Science, European History, World History, and Economics. The most marginal subjects—meaning those that experience the most changes in both directions—are Psychology, World History, and Computer Science.

3.5.2 Reduced Form Effect of AP Course Availability on College Selectivity and Graduation

To identify the causal effect of course availability on the probability of enrolling in a competitive college, graduating within four years of leaving high school, and graduating within six years, I estimate Equation 3.1 on the sample of seniors in Michigan public high schools. The reduced form estimates appear in Table 3.2. On average, there is no effect of an additional AP course offering on any of the outcomes; all of treatment coefficients are close to zero, none are significant, and they are estimated precisely. For example, the effect on enrolling in a competitive college is 0.2 percentage points, with a standard error of 0.2 percentage points. The effect on four-year BA attainment is a statistically insignificant 0.1 percentage points, and the effect on six-year degree attainment is precisely zero (to three decimal points).

As outlined in Section 3.3, there is reason to believe that the effects of AP vary by student type. In particular, more academically prepared students might be more likely to reap the benefits of a college-level curriculum. Less prepared students might fall further behind their peers if they're pushed into courses beyond their preparation level, or might be harmed by the diversion of resources towards AP students and teachers. Structural inequities may prevent low-income and underrepresented minority students from accessing advanced classes, as well as contribute to their lower levels of academic preparation. To test this, I estimate reduced form effects of AP course availability, this time interacting the treatment variables with an income proxy (free or reduced price lunch eligibility), student race (an indicator for underrepresented minority), and middle school math test score.

Composition and English Language and Composition, and Calculus AB and BC. (See Appendix Figure B.1.) Still, the hierarchy in Table B.1 corresponds to national and Michigan AP exam data from the College Board.

Table 3.3 displays the effects of interest from estimating equation 3.2, which gives the reduced form effect of AP availability by family income, which I proxy by eligibility for subsidized meals. Overall, the results suggest that higher-income students have improved outcomes when more AP courses are offered at their schools, while low-income students have worse outcomes. For low-income students, having an additional AP course makes them 0.4 percentage points *less* likely to enroll in a competitive college (an effect of 2 percent relative to the mean of 22.5 percent), 0.6 percentage points (13 percent) less likely to earn a bachelor's degree within four years of high school graduation, and 0.7 percentage points (6 percent) less likely to earn a degree within six years. For higher-income students, all of the effects are positive: an increase in competitive college enrollment of 0.4 percentage points (1 percent), an increase in four-year graduation of 0.3 percentage points (1 percent), and a statistically insignificant increase in six-year graduation of 0.2 percentage points. These effects are substantively small. The average cohort has 72 low-income students and 215 non-low-income students in its senior class, so these effects translate to not even one fewer low-income student and one more high-income students graduating in four years for the average school cohort.

Examining heterogeneity by race (Table 3.4) suggests a similar story . While White and Asian students seem to benefit from expanded AP course availability, underrepresented minority students (Black, Hispanic, and Native students)—who have lower rates of selective college attendance and completion to begin with—have worse outcomes. The effects on all three outcomes are negative for URM students: -0.3, -0.7, and -0.6 percentage points on competitive college enrollment, four-year graduation, and six-year graduation, respectively (though the effect on selective enrollment is not statistically significant). For White and Asian students, the effects are small and positive: 0.3 percentage points for competitive enrollment ($p<0.1$), 0.2 percentage points for four-year graduation ($p<0.05$), and 0.1 percentage points for six-year graduation (not statistically significant).

I also examine heterogeneity by prior academic achievement, which I measure using standardized scores on the state middle school math test. Table 3.5 summarizes the reduced form effect of AP course availability by prior test scores. While a student with average middle school test scores sees no improvement in outcomes from an additional course offering, higher-achieving students do. The interaction terms with test score for the two graduation outcomes are positive and significant at the $\alpha = 0.05$ level. A one standard deviation increase in test score increases the size of the AP course effect on four-year graduation by 0.6 percentage points and on six-year graduation by 0.2 percentage points. For a student with a math score one standard deviation above the mean, the additional available AP increases their chance of on-time college graduation by 0.5 percentage points.

Together, the heterogeneity analyses imply that expanding AP course availability widens gaps in college outcomes by income, race, and prior achievement. These findings could be consistent with several plausible stories. It is important to distinguish whether the negative reduced form effects for disadvantaged students are driven by direct negative consequences of taking more AP courses, or by a diversion of resources going to AP courses that they aren't able to take advantage of. I investigate this further in the next section.

3.5.3 Instrumental Variables Approach and First Stage Results

I turn now to the instrumental variables approach described in Section 3.4.1. Under the assumption that AP course availability affects students' outcomes only through its effect on their AP course-taking, the IV approach lets us estimate the casual effect of AP course- and test-taking, for students induced to take another AP course or test when it becomes available. This assumption would be violated if there are other, indirect effects of more AP courses at a school, such as positive peer effects or negative effects of diverted resources; the direction of the bias is not obvious.

The first stage results are by themselves useful for understanding whether and which students take advantage of additional AP courses when they are offered, and in interpreting the reduced form effects, so I present those results separately. When I estimate the first stage effect of AP course availability on the number of AP courses and exams that students take in Table 3.6, the point estimates suggest that an additional AP course offering increases the number of AP courses the average student takes by 0.031, and has no detectable effect on the number of exams. To put these magnitudes in context, the average student in my sample takes 0.78 AP courses, so this represents an increase in course-taking of 4 percent. Put differently, the average senior class has around 300 students, so these numbers translate into 9 additional students taking an AP course. This helps explain why the reduced form effects in Table 3.2 are close to zero and not statistically significant: even if AP courses improve outcomes for the students who take them, very few students take an additional AP course when it becomes available. Note that the first-stage F statistics are small and below conventional thresholds, suggesting the instrument of course availability is not sufficiently strong to estimate effects using an IV model. For this reason, I do not report IV estimates.

It is particularly important to understand the negative reduced form effects for more disadvantaged students. If the more disadvantaged students take AP courses when they become available, this would suggest a direct negative effect, perhaps from being pushed into college-level courses that they are not ready for. If, on the other hand, the more disadvantaged students do not take AP courses when they become available, the negative reduced form effect implies negative spillovers of AP courses at the school.

Tables 3.7, 3.8, and 3.9 show first stage results estimated by family income, race, and prior academic achievement. The first stage estimates in Table 3.7 suggest that for higher-income students, an additional course offering leads to higher rates of course- and exam-taking (an additional 0.038 courses and 0.023 exams); however, for lower-income students, the first stage is essentially zero. Looking at the effects by race in Table 3.8, White and Asian students take 0.37 more AP courses and 0.22 more AP exams, on average. There is no first-stage effect on course-taking for Black, Hispanic, and Native students, and the effect on exams is actually negative (-0.028 exams) for these students. Finally, Table 3.9 indicates that higher-achieving students are more likely to take advantage of additional AP courses. While an average student increases their course-taking by 0.22 courses, a student with a test score one standard deviation above average increases their course-taking by 0.043 courses more, or 0.065 courses. In terms of exam taking, there is no first stage effect on exams for average-performing students, but a significant effect for high-performing ones. I estimate that a student with a middle school math score one standard deviation above average increases their number of AP exams by 0.060—nearly the same increase as in course-taking.

Again, the first stage F statistics are small, so I do not report IV estimates for these subgroups. However, the first stage results are telling. First, they suggest that it is the already advantaged students who take additional AP courses when they become available, widening gaps in advanced course-taking. Furthermore, the largely null first-stage results on course-taking for low-income, underrepresented minority, and lower-achieving students combined with the negative reduced form effects imply that low-income and underrepresented minority students are indirectly harmed by having more AP course offerings at their school. There are several plausible channels by which this could occur. The widening of AP course-taking gaps could reflect an increase in academic tracking, and the loss of positive peer effects could harm the students on the lower track. The disadvantaged students could also lose out if the best teachers and materials are diverted into the AP courses at the expense of non-AP courses.

3.6 Threats to Identification and Robustness Checks

Because I am not able to randomly assign schools to offer AP courses, I have to worry about whether my results are picking up a true causal effect or are driven by some spurious correlation. There are several main threats to identification. Perhaps both AP participation and gaps in college enrollment are growing over time, but the former is not causing the latter. My empirical strategy addresses this in several ways: first, I include year fixed effects to allow for a time trend in college outcomes. Second, as I show in Appendix Figure B.5,

while there is an upward trend in number of AP courses at most schools, it is by no means strictly monotonic, meaning I am identifying off of changes in AP offerings in both directions. A related issue of confounding endogeneity is that it is possible that longer-term, systematic changes to the student population and the demand for AP courses are occurring, and that these are correlated with student outcomes. This would be the case if, for example, schools offer more AP courses in order to attract higher-achieving students. I test for this type of endogeneity in two ways.

First, I follow Darolia et al. (2020), who use similar data and an identification strategy identical to mine to estimate the effect of additional high school STEM course offerings on college outcomes. They discuss the issue of "the potential for endogenous changes in course offerings within high schools over time" in a way that is highly relevant to my identification strategy and leads to a straightforward robustness check. They "hypothesize that if bias from endogenous changes within high schools is present, model replications based on data that cover a shorter time span will be less biased because there is less time for major changes."

I do a similar exercise where I partition my eight years of data into two shorter panels of four years each (2005 to 2008 and 2009 to 2012), and estimate the effects of AP separately by time period. I do this for the reduced form results on all students as well as the results by FRPL status. Table 3.10 reports the reduced form estimates (Equation 3.1), estimated separately on the two shorter time periods. The null findings for selective college enrollment and four-year graduation from Table 3.2 generally hold, with the exception of a small but significant effect (0.3 percentage points) of course availability on four-year graduation in the 2005-2008 period.

The reduced form results by income status also hold up when estimated on the two shorter panels (Table 3.11). It is still the case that lower income students benefit less—possibly even lose—from having additional AP courses at their school. It is interesting to note that the effects for higher-income students are smaller and less positive in the later period, which could reflect changes in the marginal course offered or in the marginal higher-income student. However, I cannot reject that the estimates are the same across the two periods so I consider this suggestive evidence only.

As a second robustness check, I directly test for positive selection of students into schools with more AP courses by estimating a version of Equation 3.1 where the left-hand-side variable is the average middle school test score of the senior class:

$$\begin{aligned}(\text{Average middle school test score})_{jt} = \alpha + \sum_{k=-2}^{2} \beta_k AP_{j,t+k} + \sum_j \delta_j \\ + \sum_{t=2003}^{2012} \lambda_t + \sum_j \tau_j \cdot t + \boldsymbol{\mu Z'_{j,t-2}} + \varepsilon_{jt}\end{aligned} \tag{3.7}$$

Note this is done at the school-year level. Positive β_k's, particularly for $k \leq 0$, would suggest that a stronger AP curriculum attracts higher-achieving students, and would cause me to worry that my findings are driven by students with better outcomes coming into schools with more AP rather than more AP causing improved outcomes. Figure 3.1 graphically depicts the estimated β_k coefficients. There is no evidence that higher-achieving students are positively selecting into schools with more AP courses.

3.7 Conclusion

Using administrative data from the state of Michigan and exploiting within-school, across-time variation in AP course offerings, I have shown that introducing more AP courses tends to help higher-income, higher-achieving, and White and Asian students at the expense of low-income, lower-achieving, and underrepresented minority students. Not only are disadvantaged students unlikely to take advantage of expanded AP course availability, they appear to have somewhat worse college outcomes as a result of larger AP programs in their schools. These two facts together suggest a story of spillover or crowdout that exacerbates existing inequalities by income and race, both in terms of access to advanced courses as well as educational outcomes.

This finding is consistent with work by historians, sociologists, and education researchers arguing that the Advanced Placement program, like many other examples of educational resources, benefits already privileged students and systematically excludes the already marginalized, thus perpetuating inequities (e.g., Schneider, 2009). For example, Rodriguez and McGuire (2019) use cross-sectional national data and instrument for AP availability with per-pupil school expenditures and find that when schools introduce additional AP courses, the Black-White gap in AP course-taking widens. They argue that their results imply opportunity hoarding by White students and families. Similarly, Solorzano and Ornelas (2002) show that Chicana and Latina students in one California district are underrepresented in AP courses, even in schools with strong AP programs. These studies, as well as the current analysis, suggest that without a concerted effort to ensure equal access for all students,

expanding AP offerings will most likely only worsen educational inequality.

Even if students were granted truly equal access to AP courses, it is not obvious that college outcome gaps would close. Recent experimental work by Conger et al. (2020) finds negative effects of experimentally introducing AP science courses, driven by less academically prepared students. My finding of smaller and possibly negative effects of AP course availability in the second half of the panel also hint that at more recent levels of access, the marginal AP course (which tends to be in a non-core subject) has little effect and may even be harmful.

The magnitudes of the effects I find are small, and imply college outcomes changing for no more than one low-income or underrepresented minority student per school cohort as a result of introducing an additional AP course. However, even if the effect is effectively null, the policy implications from the current as well as previous work are similar: putting financial and legal resources towards expanding AP access is unlikely to achieve the goal of closing gaps in educational outcomes.

Table 3.1: Sample Descriptive Statistics

	Mean	Std dev	Min	Max	N non-missing
Student level characteristics					
Female	0.51	0.50	0.00	1.00	174,395
White	0.74	0.44	0.00	1.00	174,469
Asian	0.04	0.20	0.00	1.00	174,469
Black	0.17	0.38	0.00	1.00	174,469
Hispanic	0.03	0.18	0.00	1.00	174,469
Native	0.01	0.09	0.00	1.00	174,469
Eligible for free or reduced-price lunch	0.24	0.43	0.00	1.00	174,469
Middle school math test score (std.)	0.26	1.01	-6.36	7.19	155,377
AP courses available junior & senior year	9.79	4.58	0.00	20.00	174,469
AP courses taken junior & senior year	0.78	1.37	0.00	11.00	174,469
AP tests taken	0.73	1.59	0.00	23.00	136,285
School-cohort level characteristics					
Average middle school math test score	0.08	0.43	-1.04	1.15	687
School enrollment	1377	500	0	2519	689
Pupil-to-teacher ratio	21.41	10.97	0.00	245.50	686
Per pupil instructional spending	6360	1783	3477	43080	686
Local unemployment rate	8.87	4.61	1.74	25.50	689
AP courses available year t and $t-1$	8.56	4.64	0.00	20.00	689

Notes: "Native" includes American Indian, Alaska Native, Native Hawaiian, and other Pacific Islander students. Middle school math test score is measured as a standardized scale score. I use eighth grade test score if available and seventh grade score if not. School-year characteristics are all measured in year $t-2$, except for AP course availability. AP course availability is measured as the number of unique AP subjects offered over two years; if a subject is offered both years, it is counted once.

Table 3.2: Reduced Form Effect of AP Course Availability on College Outcomes

	(1) Enrolled in competitive+ college	(2) Earned BA degree in 4 years	(3) Earned BA degree in 6 years
# of AP courses available at school in junior and senior year	0.002 (0.002)	0.001 (0.001)	0.000 (0.001)
Mean of outcome variable	[0.391]	[0.156]	[0.318]
Observations	174,469	174,469	174,469
Cohorts	2005-2012	2005-2012	2005-2012

Notes: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$. Regressions include school fixed effects, year fixed effects, school-specific linear time trends, student-level controls (race, gender, free or reduced price lunch status, and middle school standardized math test score), and time-varying school-level controls (average middle school test score, school enrollment, pupil:teacher ratio, per student instructional spending, and local unemployment, all measured in the student's sophomore year). Robust standard errors clustered at the school level in parentheses.

Table 3.3: Reduced Form Effect of AP Course Availability on College Outcomes, by Family Income

	(1) Enrolled in competitive+ college	(2) Earned BA degree in 4 years	(3) Earned BA degree in 6 years
Effect of one additional available AP course for:			
Low-income students	-0.004**	-0.006***	-0.007***
	(0.002)	(0.001)	(0.002)
	[0.225]	[0.045]	[0.130]
Non-low-income students	0.004**	0.003***	0.002
	(0.002)	(0.001)	(0.001)
	[0.444]	[0.193]	[0.379]
Observations	174,469	174,469	174,469
Cohorts	2005-2012	2005-2012	2005-2012

Notes: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$. Low-income status is proxied by eligibility for free or reduced-price lunch (FRPL). Effects by income are estimated using a single equation, where course availability is interacted with FRPL status. Regressions include school fixed effects, year fixed effects, school-specific linear time trends, student-level controls (race, gender, free or reduced price lunch status, and middle school standardized math test score), and time-varying school-level controls (average middle school test score, school enrollment, pupil:teacher ratio, per student instructional spending, and local unemployment, all measured in the student's sophomore year). Robust standard errors clustered at the school level in parentheses. Means of the outcome variables are in brackets.

Table 3.4: Reduced Form Effect of AP Course Availability on College Outcomes, by Race and Ethnicity

	(1) Enrolled in competitive+ college	(2) Earned BA degree in 4 years	(3) Earned BA degree in 6 years
Effect of one additional available AP course for:			
Black, Hispanic, & Native students	-0.003	-0.007***	-0.006***
	(0.002)	(0.001)	(0.002)
	[0.281]	[0.054]	[0.145]
White & Asian students	0.003*	0.002**	0.001
	(0.002)	(0.001)	(0.001)
	[0.421]	[0.185]	[0.365]
Observations	174,469	174,469	174,469
Cohorts	2005-2012	2005-2012	2005-2012

Notes: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$. Effects by race are estimated using a single equation, where course availability is interacted with an indicator for being Black, Hispanic, American Indian, or Native Hawaiian or Pacific Islander. Regressions include school fixed effects, year fixed effects, school-specific linear time trends, student-level controls (race, gender, free or reduced price lunch status, and middle school standardized math test score), and time-varying school-level controls (average middle school test score, school enrollment, pupil:teacher ratio, per student instructional spending, and local unemployment, all measured in the student's sophomore year). Robust standard errors clustered at the school level in parentheses. Means of the outcome variables are in brackets.

Table 3.5: Reduced Form Effect of AP Course Availability on College Outcomes, by Academic Preparation

	(1) Enrolled in competitive+ college	(2) Earned BA degree in 4 years	(3) Earned BA degree in 6 years
# AP courses available at school in junior & senior year	0.002	-0.001	-0.001
	(0.002)	(0.001)	(0.001)
Middle school math test score	0.189***	0.047***	0.130***
	(0.009)	(0.008)	(0.011)
# of AP courses available * math score	-0.001	0.006***	0.002***
	(0.001)	(0.001)	(0.001)
Observations	155,377	155,377	155,377
Cohorts	2005-2012	2005-2012	2005-2012
Effect for students with average math score	0.002	-0.001	-0.001
Outcome mean, math score in (-0.25, 0.25)	[0.344]	[0.095]	[0.262]
Effect for students 1 sd above average	0.001	0.005***	0.002
Outcome mean, math score in (0.75, 1.25)	[0.588]	[0.257]	[0.491]

Notes: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$. Middle school math test score is measured as a standardized scale score. I use eighth grade test score if available and seventh grade score if not. Students missing a test score are not included in this analysis. Effects by academic preparation are estimated using a single equation, where course availability is interacted with a the continuous measure of test score. Regressions include school fixed effects, year fixed effects, school-specific linear time trends, student-level controls (race, gender, free or reduced price lunch status, and middle school standardized math test score), and time-varying school-level controls (average middle school test score, school enrollment, pupil:teacher ratio, per student instructional spending, and local unemployment, all measured in the student's sophomore year). Robust standard errors clustered at the school level in parentheses.

Table 3.6: First Stage Effect of AP Course Availability on AP Course- and Exam-Taking

	(1) First stage: # AP courses taken	(2) First stage: # AP exams taken
# of AP courses available at school in junior and senior year	0.031***	0.013
	(0.010)	(0.008)
Mean of outcome variable	[0.780]	[0.731]
Kleibergen-Paap Wald F statistic	9.28	2.59
Observations	174,469	136,285
Cohorts	2005-2012	2007-2012

Notes: $*p < 0.1$, $**p < 0.05$, $***p < 0.01$. Regressions include school fixed effects, year fixed effects, school-specific linear time trends, student-level controls (race, gender, free or reduced price lunch status, and middle school standardized math test score), and time-varying school-level controls (average middle school test score, urbanicity, size of senior class, pupil:teacher ratio, per student instructional spending, and local unemployment, all measured in the student's sophomore year). Robust standard errors clustered at the school level in parentheses.

Table 3.7: First Stage Effect of AP Course Availability on AP Course- and Exam-Taking, by Family Income

	(1) First stage: # AP courses taken	(2) First stage: # AP exams taken
Effect of one additional available AP course for:		
Low-income students	0.008	-0.015*
	(0.010)	(0.009)
	[0.359]	[0.292]
Non-low-income students	0.038***	0.023***
	(0.010)	(0.007)
	[0.916]	[0.888]
Kleibergen-Paap Wald F statistic	4.28	0.58
Observations	174,469	136,285
Cohorts	2005-2012	2007-2012

Notes: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$. Low-income status is proxied by eligibility for free or reduced-price lunch (FRPL). Effects by income are estimated using a single equation, where course availability is interacted with FRPL status. Regressions include school fixed effects, year fixed effects, school-specific linear time trends, student-level controls (race, gender, free or reduced price lunch status, and middle school standardized math test score), and time-varying school-level controls (average middle school test score, school enrollment, pupil:teacher ratio, per student instructional spending, and local unemployment, all measured in the student's sophomore year). Robust standard errors clustered at the school level in parentheses. Means of the course- and exam-taking variables are in brackets.

Table 3.8: First Stage Effect of AP Course Availability on AP Course- and Exam-Taking, by Race and Ethnicity

	(1) First stage: # AP courses taken	(2) First stage: # AP exams taken
Effect of one additional available AP course for:		
Black, Hispanic, & Native students	0.001	-0.028***
	(0.011)	(0.011)
	[0.357]	[0.245]
White & Asian students	0.037***	0.022***
	(0.010)	(0.007)
	[0.896]	[0.865]
Kleibergen-Paap Wald F statistic	4.20	0.99
Observations	174,469	136,285
Cohorts	2005-2012	2007-2012

Notes: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$. Effects by race are estimated using a single equation, where course availability is interacted with an indicator for being Black, Hispanic, American Indian, or Native Hawaiian or Pacific Islander. Regressions include school fixed effects, year fixed effects, school-specific linear time trends, student-level controls (race, gender, free or reduced price lunch status, and middle school standardized math test score), and time-varying school-level controls (average middle school test score, school enrollment, pupil:teacher ratio, per student instructional spending, and local unemployment, all measured in the student's sophomore year). Robust standard errors clustered at the school level in parentheses. Means of the course- and exam-taking variables are in brackets.

Table 3.9: First Stage Effect of AP Course Availability on AP Course- and Exam-Taking, by Prior Achievement

	(1) First stage: # AP courses taken	(2) First stage: # AP exams taken
# AP courses available at school in junior & senior year	0.022*	-0.005
	(0.011)	(0.008)
Middle school math test score	0.240***	0.078
	(0.055)	(0.067)
# of AP courses available * math score	0.043***	0.065***
	(0.005)	(0.006)
Kleibergen-Paap Wald F statistic	3.86	0.32
Observations	155,377	123,669
Cohorts	2005-2012	2007-2012
Effect for students with average math score	0.022*	-0.005
Outcome mean, math score in (-0.25, 0.25)	[0.424]	[0.303]
Effect for students 1 sd above average	0.065***	0.060***
Outcome mean, math score in (0.75, 1.25)	[1.239]	[1.129]

Notes: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$. Middle school math test score is measured as a standardized scale score. I use eighth grade test score if available and seventh grade score if not. Students missing a test score are not included in this analysis. Regressions include school fixed effects, year fixed effects, school-specific linear time trends, student-level controls (race, gender, free or reduced price lunch status, and middle school standardized math test score), and time-varying school-level controls (average middle school test score, school enrollment, pupil:teacher ratio, per student instructional spending, and local unemployment, all measured in the student's sophomore year). Robust standard errors clustered at the school level in parentheses. Means of the course- and exam-taking variables are in brackets.

Table 3.10: Reduced Form Effect of AP Course Availability on College Outcomes, by Shorter Time Periods

	2005-2008		
	(1) Enrolled in competitive+ college	(2) Earned BA degree in 4 years	(3) Earned BA degree in 6 years
# of AP courses available at school in junior and senior year	0.004	0.003**	0.000
	(0.003)	(0.001)	(0.002)
Observations	83,560	83,560	83,560
	2009-2012		
	(4) Enrolled in competitive+ college	(5) Earned BA degree in 4 years	(6) Earned BA degree in 6 years
# of AP courses available at school in junior and senior year	-0.001	-0.000	-0.002
	(0.002)	(0.002)	(0.002)
Observations	90,909	90,909	90,909

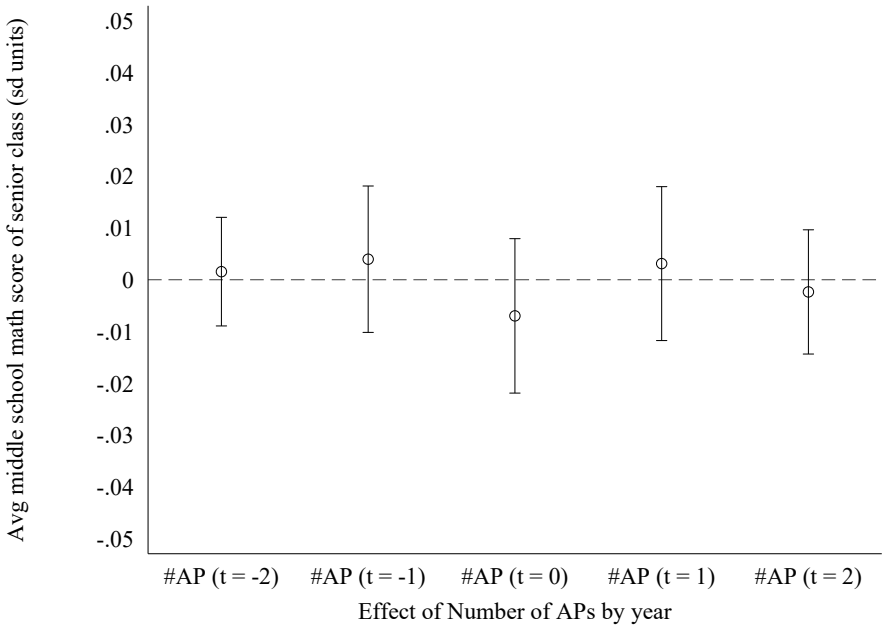
Notes: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$. Regressions include school fixed effects, year fixed effects, school-specific linear time trends, student-level controls (race, gender, free or reduced price lunch status, and middle school standardized math test score), and time-varying school-level controls (average middle school test score, school enrollment, pupil:teacher ratio, per student instructional spending, and local unemployment, all measured in the student's sophomore year). Robust standard errors clustered at the school level in parentheses.

Table 3.11: Reduced Form Effect of AP Course Availability on College Outcomes by Family Income, by Shorter Time Periods

	Earlier panel: 2005-2008		
	(1) Enrolled in competitive+ college	(2) Earned BA degree in 4 years	(3) Earned BA degree in 6 years
Effect of one additional available AP course for:			
Low-income students	-0.003	-0.003*	-0.008***
	(0.003)	(0.002)	(0.002)
Non-low-income students	0.005*	0.005***	0.002
	(0.003)	(0.001)	(0.002)
Observations	83,560	83,560	83,560
	Later panel: 2009-2012		
	(4) Enrolled in competitive+ college	(5) Earned BA degree in 4 years	(6) Earned BA in 6 years
Effect of one additional available AP course for:			
Low-income students	-0.006***	-0.006***	-0.008***
	(0.002)	(0.002)	(0.002)
Non-low-income students	0.001	0.002	-0.000
	(0.002)	(0.002)	(0.002)
Observations	90,909	90,909	90,909

Notes: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$. Low-income status is proxied by eligibility for free or reduced-price lunch (FRPL). Effects by income are estimated using a single equation, where course availability is interacted with FRPL status. Regressions include school fixed effects, year fixed effects, school-specific linear time trends, student-level controls (race, gender, free or reduced price lunch status, and middle school standardized math test score), and time-varying school-level controls (average middle school test score, school enrollment, pupil:teacher ratio, per student instructional spending, and local unemployment, all measured in the student's sophomore year). Robust standard errors clustered at the school level in parentheses.

Figure 3.1: Test for Selection: Effect of Number of AP Courses Available on Average Middle School Math Test Scores of Senior Class



APPENDICES

APPENDIX A

Appendix to College Field Specialization and Beliefs about Relative Performance: An Experimental Intervention to Understand Gender Gaps in STEM

Figure A.1: Sample Intervention Message: Information-Only Treatment

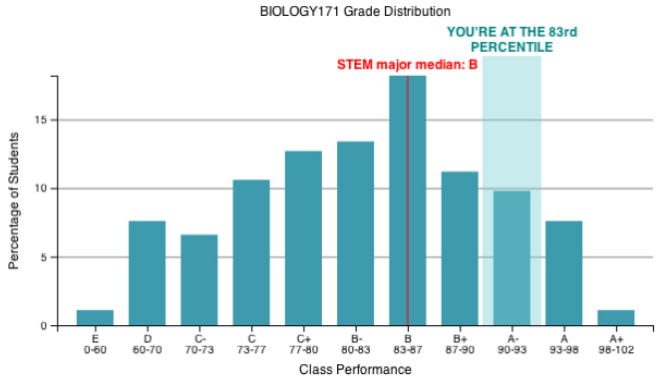
Your Bio 171 grade | And your major

[redacted],

A lot of people think they have to get *perfect* grades in the required classes to major in something. We're here to tell you: **it's not true**.

HERE'S HOW YOU'RE DOING.

This chart shows the distribution of scores for students in BIOLOGY 171 (as of November 11, 2019).



- Your score is 90.8.
- You're doing as well as or better than 83% of your classmates.

HERE'S HOW GRADES OFTEN LOOK.

The typical median grade for BIOLOGY 171 is:

- **B** for all students in BIOLOGY 171
- **B+** for BIOLOGY 171 students who declare a biology major
- **B** for BIOLOGY 171 students who declare a major in math, science, engineering, or economics

Surprised? We were, too, and we wanted to share the news with you.

In case you forgot, median means half the people are below it and half are above it.

AS YOU PLAN YOUR SCHEDULE...

A degree in biology — or another quantitative field like math, science, engineering, or economics — can open many doors.

If you want to learn more about these majors, consider scheduling an advising appointment:

- Biology
- LSA natural science major
- Computer Science
- Engineering
- Mathematics
- Economics

You can view course options for Winter 2020 here.

~ The ECoach Team

Figure A.2: Sample Intervention Message: Information-Plus-Encouragement Treatment

Your Bio 171 grade | And your major

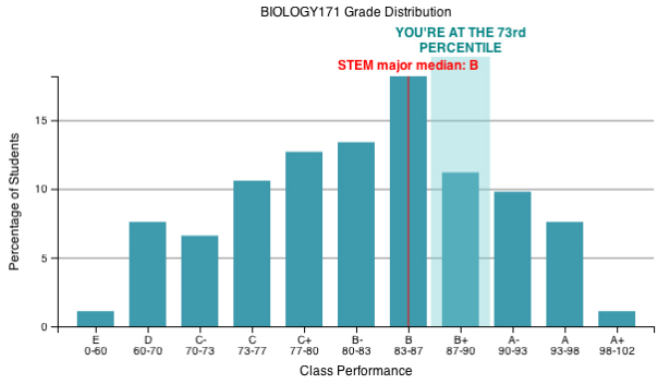
[redacted]

A lot of people think they have to get *perfect* grades in the required classes to major in something. We're here to tell you: **it's not true**.

In fact, **you're doing great** and we'd like YOU to **consider a major** in biology — or another quantitative field like math, science, engineering, or economics.

YOU'RE PERFORMING LIKE A STEM MAJOR!

This chart shows the distribution of scores for students in BIOLOGY 171 (as of November 11, 2019).



Congratulations! Your scores mean you're doing better than most students who go on to major in STEM.

- With your strong performance, your instructors hope you'll **consider a major** in biology, or another quantitative field like math, science, engineering, or economics.
- Your score is 87.9.
- You're doing as well as or better than 73% of your classmates.

HERE'S HOW GRADES OFTEN LOOK.

The typical median grade for BIOLOGY 171 is:

- **B** for all students in BIOLOGY 171
- **B+** for BIOLOGY 171 students who declare a biology major
- **B** for BIOLOGY 171 students who declare a major in math, science, engineering, or economics

Surprised? We were, too, and we wanted to share the news with you.

In case you forgot, median means half the people are below it and half are above it.

AS YOU PLAN YOUR SCHEDULE...

A degree in biology — or another quantitative field like math, science, engineering, or economics — can open many doors.

We hope you'll learn more about these majors. One way is to schedule an advising appointment:

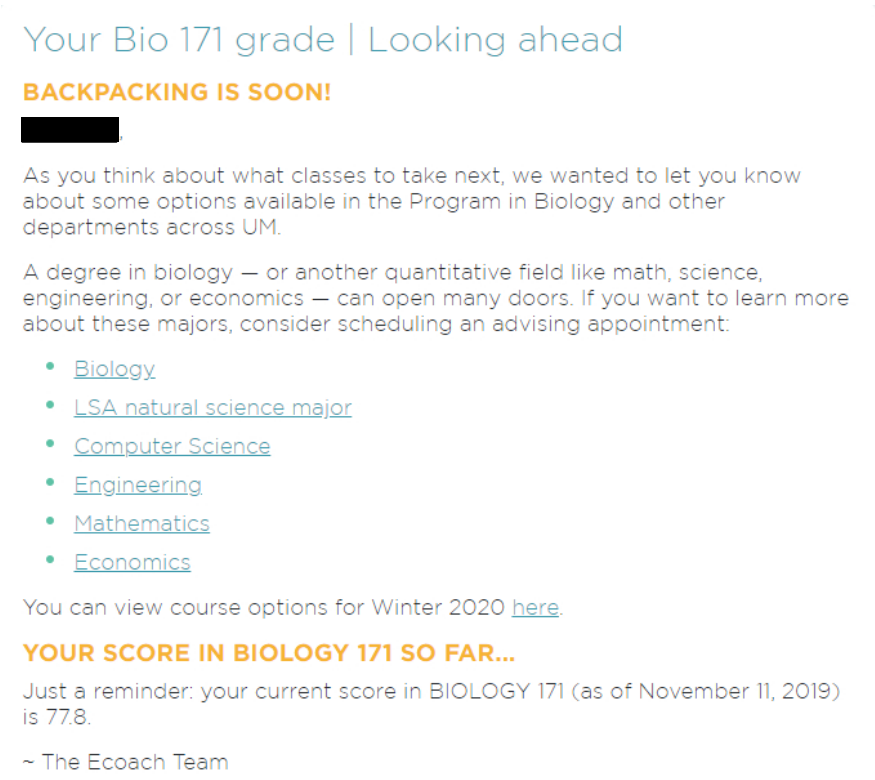
- Biology
- LSA natural science major
- Computer Science
- Engineering
- Mathematics
- Economics

You can view course options for Winter 2020 here.

Congrats again — keep up the good work!

~ The ECoach Team

Figure A.3: Sample Intervention Message: Control Group



Your Bio 171 grade | Looking ahead

BACKPACKING IS SOON!

[redacted]

As you think about what classes to take next, we wanted to let you know about some options available in the Program in Biology and other departments across UM.

A degree in biology — or another quantitative field like math, science, engineering, or economics — can open many doors. If you want to learn more about these majors, consider scheduling an advising appointment:

- Biology
- LSA natural science major
- Computer Science
- Engineering
- Mathematics
- Economics

You can view course options for Winter 2020 here.

YOUR SCORE IN BIOLOGY 171 SO FAR...

Just a reminder: your current score in BIOLOGY 171 (as of November 11, 2019) is 77.8.

~ The Ecoach Team

Table A.1: Balance by Assignment to Information-Only and Information-Plus-Encouragement Treatment, Above-Median Students Only

	Control	Info-only	Info + encour.	p-value
Female	0.461	0.459	0.461	
Class standing (omitted: senior)				
First year	0.418	0.420	0.404	0.728
Sophomore	0.419	0.411	0.428	0.764
Junior	0.126	0.125	0.127	0.993
Race/ethnicity (omitted: American Indian or multiple race/ethnicities)				
White	0.566	0.527	0.555	0.181
Hispanic	0.041	0.055	0.044	0.279
Asian	0.319	0.343	0.330	0.495
Black	0.013	0.007	0.014	0.248
Declared major (omitted: other)				
Undeclared	0.545	0.541	0.539	0.964
Engineering	0.260	0.255	0.266	0.708
Math, science, or economics	0.104	0.112	0.091	0.315
Academic and demographic characteristics				
In-state	0.480	0.460	0.490	0.409
Prior college GPA	3.612	3.610	3.626	0.827
Math placement score (std)	0.330	0.365	21.002	0.552
ACT English	33.380	33.289	33.533	0.379
ACT Math	32.336	32.279	32.375	0.810
ACT Reading	32.696	32.310	32.740	0.052
ACT Science	32.193	32.102	32.160	0.897
SAT Math	737.577	738.541	734.895	0.301
SAT Verbal	661.075	658.928	660.807	0.905
HS GPA	3.916	3.916	3.912	0.614
Took calculus in HS	0.873	0.882	0.858	0.308
Max parental education (omitted: less than high school)				
High school	0.042	0.055	0.040	0.254
Some college	0.038	0.029	0.037	0.525
Bachelor's	0.242	0.221	0.248	0.374
Grad or professional degree	0.669	0.683	0.663	0.623
Family Income (omitted: less than $50,000)				
$50,000-100,000	0.158	0.170	0.166	0.805
Above $100,000	0.731	0.704	0.716	0.505
Total N	940	943	940	2,823

Notes: Sample limited to above-median students; only above-median students were eligible for the information-plus-encouragement treatment. P-values based on a joint test of differences in the characteristic by treatment status, controlling for strata. I also test for differences in missingness rates on all variables and find none.

Table A.2: Balance by Assignment to Treatment, by Gender

	Men			Women		
	Control	Treat	p-value	Control	Treat	p-value
Class standing (omitted: senior)						
First year	0.446	0.407	0.077	0.419	0.428	0.688
Sophomore	0.370	0.405	0.237	0.406	0.401	0.711
Junior	0.135	0.136	0.813	0.129	0.128	0.934
Race/ethnicity (omitted: American Indian or multiple race/ethnicities)						
White	0.560	0.543	0.476	0.556	0.544	0.380
Hispanic	0.078	0.072	0.876	0.062	0.064	0.301
Asian	0.258	0.300	0.203	0.248	0.277	0.486
Black	0.025	0.018	0.664	0.052	0.033	0.200
Declared major (omitted: other)						
Undeclared	0.487	0.477	0.947	0.638	0.650	0.417
Engineering	0.305	0.314	0.842	0.153	0.149	0.388
Math, science, or economics	0.103	0.102	0.767	0.086	0.086	0.739
Academic and demographic characteristics						
In-state	0.514	0.506	0.688	0.534	0.536	0.368
Prior college GPA	3.296	3.368	0.806	3.444	3.483	0.362
Math placement score (std)	0.080	0.242	0.077	-0.251	-0.146	0.560
ACT English	32.439	32.532	0.285	32.217	32.691	0.387
ACT Math	31.851	32.122	0.641	29.848	30.386	0.661
ACT Reading	31.975	31.761	0.026	31.981	31.934	0.101
ACT Science	31.629	31.810	0.463	30.124	30.405	0.459
SAT Math	717.445	729.825	0.128	690.168	694.202	0.019
SAT Verbal	646.050	653.934	0.289	637.603	639.435	0.155
HS GPA	3.871	3.880	0.685	3.895	3.901	0.648
Took calculus in HS	0.832	0.867	0.097	0.796	0.806	0.651
Max parental education (omitted: less than high school)						
High school	0.069	0.062	0.998	0.072	0.079	0.125
Some college	0.052	0.043	0.583	0.077	0.061	0.529
Bachelor's	0.242	0.237	0.973	0.265	0.245	0.275
Grad or professional degree	0.612	0.639	0.646	0.561	0.593	0.785
Family Income (omitted: less than $50,000)						
$50,000-100,000	0.175	0.185	0.308	0.190	0.195	0.462
Above $100,000	0.658	0.664	0.392	0.588	0.619	0.990
P-value on F test of all X's		0.8306			0.7071	
Total N	1,240	1,753	2,993	1,142	1,580	2,722

Notes: "Treat" column includes students receiving either treatment arm. P-values based on a regression of the characteristic on treatment status, controlling for strata. I also test for differences in missingness rates on all variables and find none. F-test tests for joint significance of all listed characteristics as well as missingness rates in predicting treatment, controlling for strata.

Table A.3: Study Sample and Gender Breakdown by Course

Course (for study)	Number of students	Proportion of sample	Course proportion women
Biology	566	0.099	0.654
Chemistry	1,127	0.197	0.531
Economics	825	0.144	0.461
Computer Science	882	0.154	0.376
Engineering	453	0.079	0.305
Physics	327	0.057	0.269
Statistics	1,535	0.269	0.531
Total	5,715	1.000	0.476
In multiple courses	855	0.150	

Notes: Students in multiple courses are assigned to a single course, chosen randomly, for purposes of the study, so that the proportions across study courses sum to 1. Course proportion women measures the proportion of students in the sample for each course who are women.

Table A.4: Intervention Message View Rate by Student Characteristics, Treated Students

Characteristic	Viewed message coef.	Characteristic	Viewed message coef.
Female	0.045**	*Declared major (omitted: other)*	
	(0.021)	Undeclared	-0.044**
Above course median	0.034*		(0.020)
	(0.020)	Engineering	-0.056*
Female*above median	0.008		(0.030)
	(0.026)	Math, science, or econ	-0.016
Course (omitted: Chemistry)			(0.028)
Biology	0.145***	*Acad. and demog. characteristics*	
	(0.027)	In state	-0.015
Econ (section 1)	0.108***		(0.015)
	(0.030)	Prior college GPA	0.081***
Econ (section 2)	0.116***		(0.025)
	(0.033)	College GPA missing	0.360***
Computer Science	0.162***		(0.090)
	(0.026)	Math placement score	0.002
Engineering	0.144***		(0.002)
	(0.031)	Placement score missing	0.046
Physics	0.129***		(0.058)
	(0.033)	ACT English	-0.005
Statistics	0.167***		(0.003)
	(0.024)	ACT math	0.003
Class standing (omitted: senior)			(0.003)
First year	0.034	ACT reading	-0.003
	(0.040)		(0.003)
Sophomore	0.039	ACT science	0.001
	(0.036)		(0.003)
Junior	0.017	ACT missing	-0.186*
	(0.037)		(0.106)
Race/ethnicity (omitted: other/multiple)		SAT math	-0.000
White	0.026		(0.000)
	(0.027)	SAT verbal	-0.000*
Hispanic	0.008		(0.000)
	(0.037)	SAT missing	-0.249**
Asian	0.016		(0.123)
	(0.029)	HS GPA	-0.009
Black	0.095**		(0.062)
	(0.046)	HS GPA missing	-0.016
Race/ethnicity missing	-0.039		(0.243)
	(0.050)	Took calculus in HS	0.008
			(0.020)
		HS calculus missing	-0.014
			(0.032)

Continued on next page

Table A.4 – *Continued from previous page*

Characteristic	Viewed message coef.
Max parent ed (omitted: less than HS)	
High school	-0.045
	(0.050)
Some college	-0.048
	(0.052)
Bachelor's	-0.023
	(0.047)
Grad or professional degree	-0.049
	(0.046)
Parent ed missing	-0.061
	(0.077)
Family income (omitted: <$50,000)	
$50,000-100,000	-0.011
	(0.026)
Above $100,000	0.006
	(0.023)
Family income missing	0.003
	(0.025)
N	3,333

Notes: $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$ Table shows coefficients and robust standard errors from a regression where the dependent variable is an indicator for viewing the intervention message. Sample limited to students assigned to treatment.

Table A.5: Survey Response Rates

	Response rate	Number of responses
Pre-intervention survey		
Overall response	0.746	4,266
Item-level response		
Belief about own performance	0.641	3,664
Belief about STEM major performance	0.685	3,915
Intended major	0.698	3,988
Post-intervention survey		
Overall response	0.487	2,784
Item-level response		
Belief about own performance	0.413	2,358
Belief about STEM major performance	0.461	2,632
Intended major	0.466	2,662
STEM interest index	0.462	2,639
General interest in STEM	0.460	2,631
Intent to seek STEM advising	0.461	2,632
Intent to take STEM courses	0.462	2,638
STEM success index	0.470	2,687
Grades good enough for STEM	0.465	2,655
Self-efficacy scale	0.464	2,651
STEM identity scale	0.461	2,636

Table A.6: Post-Intervention Survey Response by Student Characteristics, Full Sample

Characteristic	Viewed message coef.	Characteristic	Viewed message coef.
Female	0.071***	*Declared major (omitted: other)*	
	(0.017)	Undeclared	0.006
Above course median	0.070***		(0.019)
	(0.017)	Engineering	0.080***
Female*above median	-0.022		(0.025)
	(0.022)	Math, science, or econ	0.031
Course (omitted: Econ section 1)			(0.027)
Biology	0.561***	*Acad. and demog. characteristics*	
	(0.024)	In state	0.009
Chemistry	0.017		(0.012)
	(0.017)	Prior college GPA	0.109***
Computer Science	0.485***		(0.020)
	(0.022)	College GPA missing	0.418***
Engineering	0.642***		(0.071)
	(0.027)	Math placement score	0.002
Physics	0.086***		(0.002)
	(0.027)	Placement score missing	-0.007
Statistics	0.641***		(0.048)
	(0.017)	ACT English	0.001
Econ (section 2)	0.610***		(0.003)
	(0.028)	ACT math	-0.001
Class standing (omitted: senior)			(0.003)
First year	0.080**	ACT reading	0.000
	(0.035)		(0.003)
Sophomore	0.086***	ACT science	-0.005*
	(0.031)		(0.003)
Junior	0.023	ACT missing	-0.168*
	(0.031)		(0.093)
Race/ethnicity (omitted: other/multiple)		SAT math	-0.000
White	0.007		(0.000)
	(0.022)	SAT verbal	-0.000***
Hispanic	0.008		(0.000)
	(0.030)	SAT missing	-0.295***
Asian	0.067***		(0.104)
	(0.024)	HS GPA	0.123**
Black	-0.032		(0.053)
	(0.039)	HS GPA missing	0.479**
Race/ethnicity missing	0.052		(0.207)
	(0.039)	Took calculus in HS	-0.001
			(0.017)
		HS calculus missing	-0.016
			(0.026)

Continued on next page

Table A.6 – *Continued from previous page*

Characteristic	Took survey coef.
Max parent ed (omitted: less than HS)	
High school	-0.000
	(0.044)
Some college	-0.024
	(0.046)
Bachelor's	0.011
	(0.041)
Grad or professional degree	-0.007
	(0.041)
Parent ed missing	0.027
	(0.064)
Family income (omitted: < $50,000)	
$50,000-100,000	0.013
	(0.022)
Above $100,000	0.026
	(0.020)
Family income missing	0.047**
	(0.022)
N	5,715

Notes: $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$ Table shows coefficients and standard errors from a regression where the dependent variable is an indicator for response to the end of term survey.

Table A.7: Balance by Assignment to Treatment, Post-Intervention Survey Respondents

	Control mean	Treatment mean	p-value	N non missing
Female	0.517	0.506		2,784
Class standing (omitted: senior)				
First year	0.411	0.392	0.308	2,784
Sophomore	0.417	0.428	0.900	
Junior	0.129	0.136	0.344	
Race/ethnicity (omitted: American Indian or multiple race/ethnicities)				
White	0.533	0.535	0.916	2,698
Hispanic	0.061	0.063	0.190	
Asian	0.304	0.317	0.641	
Black	0.030	0.019	0.257	
Declared major (omitted: other)				
Undeclared	0.601	0.574	0.258	2,784
Engineering	0.201	0.209	0.299	
Math, science, or economics	0.095	0.104	0.504	
Academic and demographic characteristics				
In-state	0.506	0.517	0.292	2,784
Prior college GPA	3.441	3.475	0.217	1,172
Math placement score (std.)	-0.025	0.107	0.868	2,676
ACT English	32.527	32.718	0.493	1,567
ACT Math	30.926	31.374	0.811	1,567
ACT Reading	32.085	31.863	0.007	1,567
ACT Science	30.881	31.118	0.363	1,567
SAT Math	708.241	716.954	0.245	1,623
SAT Verbal	640.119	647.132	0.813	1,623
HS GPA	3.888	3.898	0.999	2,374
Took calculus in HS	0.817	0.842	0.721	2,506
Max parental education (omitted: less than high school)				
High school	0.069	0.066	0.378	2,746
Some college	0.061	0.049	0.579	
Bachelor's	0.255	0.241	0.376	
Grad or professional degree	0.593	0.624	0.636	
Family Income (omitted: less than $50,000)				
$50,000-100,000	0.192	0.185	0.959	2,096
Above $100,000	0.628	0.659	0.919	
P-value on F test of all X's		0.9532		
Total N	1,154	1,630	2,784	

Notes: Sample limited to students who responded to post-intervention survey. "Treatment" includes students receiving either treatment arm. P-values based on a regression of the characteristic on treatment status, controlling for strata. I also test for differences in missingness rates on all variables and find none. F-test tests for joint significance of all listed characteristics (except for female, which is blocked on) as well as missingness rates in predicting treatment, controlling for strata.

Table A.8: Comparison of Model-based and Randomization Inference P-values for Main Results

	Main effect		Effect for men		Effect for women		Men-women diff.	
Outcome	Model p-value	Rand. p-value	Model p-value	Rand. p-value	Model p-value	Rand. p-value	Model p-value	Rand. p-value
Absolute value percentile error	0.024	0.025	0.026	0.021	0.387	0.383	0.259	0.249
Signed percentile error	0.486	0.476	0.673	0.665	0.570	0.569	0.948	0.949
Underestimating STEM median	0.022	0.022	0.021	0.020	0.400	0.395	0.220	0.218
Overestimating STEM median	0.217	0.223	0.782	0.778	0.045	0.048	0.111	0.115
Number of STEM credits	0.056	0.053	0.033	0.032	0.573	0.566	0.303	0.300
Took any STEM courses	0.061	0.063	0.129	0.132	0.241	0.247	0.975	0.976

Notes: Each pair of p-values correspond to a single test statistic. Model-based p-values correspond to the analyses in Tables 1.3 and 1.5. Randomization-based p-values are based on 10,000 random draws from the distribution of possible treatment assignments, where treatment is assigned according to the procedure used for original randomization, and the test statistic is calculated the same way as for estimation. Randomization p-value is calculated as the proportion of simulated effects that are larger in absolute value than the observed effect.

Table A.9: Statistical Significance of Main Results, Adjusted for Multiple Hypothesis Testing

	Effect	Unadjusted p-value	FDR 1-stage q-value	FDR 2-stage q-value	FWER p-value
Beliefs outcomes					
Absolute value of percentile error					
Overall	-1.485	0.024	0.048	0.051	0.086
Men	-2.243	0.026	0.053	0.055	0.082
Women	-0.743	0.387	0.534	0.667	0.767
Difference, M vs. W		0.259	0.346	0.529	0.526
Signed percentile error					
Overall	0.592	0.486	0.486	0.321	0.485
Men	0.536	0.673	0.783	0.643	0.892
Women	0.647	0.570	0.570	0.746	0.767
Difference, M vs. W		0.948	0.949	0.529	0.950
Underestimating STEM median					
Overall	-0.033	0.022	0.048	0.051	0.086
Men	-0.052	0.021	0.053	0.055	0.082
Women	-0.016	0.400	0.534	0.667	0.767
Difference, M vs. W		0.220	0.346	0.529	0.526
Overestimating STEM median					
Overall	-0.023	0.217	0.290	0.170	0.386
Men	0.007	0.782	0.783	0.643	0.892
Women	-0.051	0.045	0.182	0.222	0.169
Difference, M vs. W		0.111	0.346	0.529	0.377
Behavior outcomes					
Number of STEM credits					
Overall	-0.182	0.056	0.061	0.065	0.096
Men	-0.276	0.033	0.066	0.071	0.057
Women	-0.079	0.573	0.574	0.932	0.567
Difference, M vs. W		0.303	0.606	1.000	0.472
Took any STEM					
Overall	-0.014	0.061	0.061	0.065	0.096
Men	-0.014	0.129	0.129	0.071	0.129
Women	-0.014	0.241	0.483	0.932	0.377
Difference, M vs. W		0.975	0.975	1.000	0.976

Notes: Each row corresponds to a single test statistic. Effects and unadjusted p-values correspond to the analyses in Tables 1.3 and 1.5. The FDR one-stage q-value is calculated using the procedure from Benjamini and Hochberg (1995). The two-stage FDR q-value is calculated using the procedure from Benjamini et al. (2006). Both adjustments control the false discovery rate (FDR). The FWER p-value is calculated using the free-step down permutation sampling (re-randomization) technique from Westfall and Young (1993) using 10,000 re-randomization iterations. This method controls the family-wise error rate (FWER). Adjustments are done within a family of tests. There are eight families of tests, defined by outcome group (beliefs outcomes or behavior outcomes) and type of test (all students, men, women, or the male-female difference).

Table A.10: Estimated Effect of Intervention on Students' Beliefs about Themselves and Other STEM Majors, Overall and by Gender, without Covariates

	Absolute value error in percentile beliefs (| Predicted - realized |)			Signed error in percentile beliefs (Predicted - realized)		
	All	Men	Women	All	Men	Women
Treatment effect	-1.509**	-2.415**	-0.626	0.543	0.414	0.669
	(0.658)	(1.006)	(0.851)	(0.845)	(1.264)	(1.126)
P-value, women vs. men			0.175			0.880
Control mean	18.981	20.331	17.646	6.361	8.471	4.276
N	2,358	1,166	1,192	2,358	1,166	1,192
	Underestimating course median for STEM majors			Overestimating course median for STEM majors		
	All	Men	Women	All	Men	Women
Treatment effect	-0.029**	-0.053**	-0.007	-0.025	0.009	-0.057**
	(0.015)	(0.022)	(0.019)	(0.018)	(0.026)	(0.026)
P-value, women vs. men			0.114			0.070
Control mean	0.206	0.257	0.159	0.46	0.368	0.545
N	2,632	1,291	1,341	2,632	1,291	1,341

Notes: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$. Treatment effects for all students estimated from a regression of the outcome on assignment to either treatment, controlling for student academic and demographic characteristics and randomization strata dummies. Treatment effects by gender estimated from a single regression of the outcome on assignment to the either treatment, female, and treatment-times-female, controlling only for randomization strata dummies. Estimates with covariates are reported in Table 1.3. Robust standard errors reported. All beliefs outcomes based on response to post-intervention survey. Realized performance measured mid-semester, at the time of intervention.

Table A.11: Estimated Effect of Intervention on Students' STEM Course-taking, Overall and by Gender, without Covariates

	Number of STEM credits one semester post intervention			Took any STEM courses one semester post intervention		
	All	Men	Women	All	Men	Women
Treatment effect	-0.201*	-0.259*	-0.137	-0.015*	-0.014	-0.015
	(0.108)	(0.148)	(0.157)	(0.008)	(0.009)	(0.012)
P-value, women vs. men			0.572			0.990
Control mean	8.507	9.476	7.454	0.91	0.936	0.881
N	5,715	2,993	2,722	5,715	2,993	2,722

Notes: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$. Treatment effects for all students estimated from a regression of the outcome on assignment to either treatment, controlling for randomization strata dummies. Treatment effects by gender estimated from a single regression of the outcome on assignment to the either treatment, female, and treatment-times-female, controlling only for randomization strata dummies. Estimates with covariates are reported in Table 1.5. Robust standard errors reported. Course-taking outcomes based on University of Michigan administrative data.

Table A.12: Estimated Effect of Intervention on Students' STEM Course-taking by Gender and Treatment Arm, Above-Median Students Only

	Number of STEM credits one semester post intervention			Took any STEM courses one semester post intervention		
	All	Men	Women	All	Men	Women
Pooled effect	-0.139	-0.271	0.015	-0.010	-0.011	-0.010
	(0.131)	(0.171)	(0.202)	(0.008)	(0.009)	(0.014)
P-value, women vs. men			0.280			0.957
Info-only effect	-0.192	-0.373*	0.021	-0.006	-0.010	-0.003
	(0.151)	(0.198)	(0.235)	(0.009)	(0.010)	(0.016)
P-value, women vs. men			0.201			0.700
Info + encouragement effect	-0.110	-0.197	-0.006	-0.015	-0.014	-0.015
	(0.151)	(0.198)	(0.231)	(0.010)	(0.011)	(0.017)
P-value, women vs. men			0.530			0.951
P-value, info vs. info+enc	0.587	0.378	0.907	0.392	0.692	0.439
Control mean	9.527	10.512	8.373	0.96	0.976	0.94
N	2,823	1,524	1,299	2,823	1,524	1,299

Notes: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$. Only above-median students were eligible for the information-plus-encouragement treatment; all below-median treated students received information only. Effect of either treatment (pooled) for above-median students estimated from a regression of outcome on an indicator for receiving either treatment, an indicator for being above the course median at time of randomization, and their interaction. To estimate effects on men and women, a full three-way interaction between treatment, female, and above-median is added. Treatment effects of the information-only and info-plus-encouragement intervention for above-median students estimated only on the sample of above-median students using the same specifications as above, but with two separate treatment indicators. All regressions control for student academic and demographic characteristics and randomization strata dummies. Robust standard errors reported. Course-taking outcomes based on University of Michigan administrative data.

Table A.13: Estimated Effect of Intervention on Students' STEM Course-taking, Limited to Survey Respondents

	Number of STEM credits one semester post intervention			Took any STEM courses one semester post intervention		
	All	Men	Women	All	Men	Women
Treatment effect	-0.120	-0.244	-0.002	-0.015	-0.014	-0.016
	(0.134)	(0.189)	(0.191)	(0.010)	(0.012)	(0.016)
P-value, women vs. men			0.368			0.907
Control mean	8.449	9.519	7.451	0.916	0.948	0.886
N	2,784	1,363	1,421	2,784	1,363	1,421

Notes: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$. Sample limited to students with a response to the post-intervention survey. Treatment effects for all students estimated from a regression of the outcome on assignment to either treatment, controlling for student academic and demographic characteristics and randomization strata dummies. Treatment effects by gender estimated from a single regression of the outcome on assignment to the either treatment, female, and treatment-times-female, controlling for student academic and demographic characteristics and randomization strata dummies. Robust standard errors reported. Course-taking outcomes based on University of Michigan administrative data.

Table A.14: Estimated Effect of Intervention on Students' Beliefs about Themselves and Other STEM Majors, Using Inverse Probability Weighting to Adjust for Survey Non-response

	Absolute value of error in percentile beliefs (|Predicted - realized|)			Signed error in percentile beliefs (Predicted - realized)		
	All	Men	Women	All	Men	Women
Treatment effect	-1.212	-2.871**	0.596	-0.192	-1.231	0.940
(inv. prob.-weighted)	(0.866)	(1.221)	(1.233)	(1.041)	(1.444)	(1.506)
P-value, women vs. men			0.048			0.300
Control mean (inv. prob.-weighted)	19.166	20.685	17.59	8.469	10.67	6.185
N	2,358	1,166	1,192	2,358	1,166	1,192
	Underestimating course median for STEM majors			Overestimating course median for STEM majors		
	All	Men	Women	All	Men	Women
Treatment effect	-0.019	-0.038	0.002	-0.012	0.017	-0.044
(inv. prob.-weighted)	(0.017)	(0.026)	(0.023)	(0.023)	(0.034)	(0.031)
P-value, women vs. men			0.243			0.187
Control mean (inv. prob.-weighted)	0.179	0.218	0.14	0.515	0.425	0.607
N	2,632	1,291	1,341	2,632	1,291	1,341

Notes: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$. Inverse probability weights (IPW) are constructed by running a logistic regression of an item response indicator on all of the characteristics listed in Table 1.1 as well as study course and an indicator for performing above the course median at the time of treatment. The IPW is equal to one over the predicted probability of response. IPW's are specific to individual survey items. Treatment effects for all students estimated from a regression of the outcome on assignment to either treatment, controlling for student academic and demographic characteristics and randomization strata dummies, weighting observations by the inverse of the predicted probability of responding to the relevant item. Treatment effects by gender estimated from a single regression of the outcome on assignment to the either treatment, female, and treatment-times-female, controlling for student academic and demographic characteristics and randomization strata dummies and weighting by the IPW. Robust standard errors reported. All beliefs outcomes based on response to post-intervention survey. Realized performance measured mid-semester, at the time of intervention. Control means are also weighted by the IPW. Unweighted estimates are shown in Table 1.3.

Table A.15: Estimated Effect of Intervention on Number of Credits in Non-STEM Subjects

	Social Science			Psychology			Business and Policy		
	All	Men	Women	All	Men	Women	All	Men	Women
Treatment effect	-0.004	-0.036	0.032	0.062	0.094	0.028	-0.036	-0.038	-0.034
	(0.045)	(0.057)	(0.070)	(0.053)	(0.061)	(0.089)	(0.029)	(0.044)	(0.038)
P-value, women vs. men			0.454			0.546			0.945
Control mean	0.717	0.657	0.783	1.006	0.594	1.454	0.339	0.396	0.277
N	5,715	2,993	2,722	5,715	2,993	2,722	5,715	2,993	2,722
	Humanities and Arts			Other					
	All	Men	Women	All	Men	Women			
Treatment effect	0.058	0.100	0.013	0.082	0.101	0.061			
	(0.079)	(0.106)	(0.119)	(0.060)	(0.073)	(0.097)			
P-value, women vs. men			0.586			0.742			
Control mean	3.219	2.874	3.593	1.157	0.894	1.443			
N	5,715	2,993	2,722	5,715	2,993	2,722			

Notes: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$. Treatment effects for all students estimated from a regression of the outcome on assignment to either treatment, controlling for student academic and demographic characteristics and randomization strata dummies. Treatment effects by gender estimated from a single regression of the outcome on assignment to the either treatment, female, and treatment-times-female, controlling for student academic and demographic characteristics and randomization strata dummies. Robust standard errors reported. Course-taking outcomes based on University of Michigan administrative data and measured in the semester following the intervention. "Social science" includes anthropology, political science, and sociology. "Humanities and arts" includes foreign languages, history, philosophy and religion, English and writing, cultural studies, and visual and performing arts. "Other" includes all other subjects. All outcomes measured as number of credits in the semester following the intervention.

Table A.16: Estimated Effects of Intervention on Students' Subjective Interest in STEM and Predicted Degree Receipt, Overall and by Gender

	Intent to major in STEM (binary)			STEM interest/intent index (std. dev. units)		
	All	Men	Women	All	Men	Women
Treatment effect	-0.019	-0.011	-0.026	-0.066**	-0.045	-0.085*
	(0.016)	(0.020)	(0.024)	(0.031)	(0.040)	(0.047)
P-value, women vs. men			0.623			0.526
Control mean	0.733	0.788	0.682	0	0.11	-0.102
N	2,662	1,302	1,360	2,639	1,289	1,350
	Predicted probabilty of obtaining a STEM degree					
	All	Men	Women			
Treatment effect	-0.006	-0.008	-0.004			
	(0.006)	(0.007)	(0.009)			
P-value, women vs. men			0.745			
Control mean	0.594	0.677	0.505			
N	5,715	2,993	2,722			

Notes: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$. Treatment effects for all students estimated from a regression of the outcome on assignment to either treatment, controlling for student academic and demographic characteristics and randomization strata dummies. Treatment effects by gender estimated from a single regression of the outcome on assignment to the either treatment, female, and treatment-times-female, controlling for student academic and demographic characteristics and randomization strata dummies. Robust standard errors reported. STEM interest and intent outcomes based on response to post-intervention survey. Predicted STEM degree is a predicted probability, based on pre-treatment characteristics and subsequent course-taking. Prediction specification estimated on a historical sample of students taking the same courses as the experimental sample.

Table A.17: Estimated Effect of Intervention by Pre-Intervention Prediction of Own Percentile (Under vs. Overpredicting)

	Absolute value percentile error	Signed percentile error	Underestimating STEM median	Overestimating STEM median	Number of STEM credits
Underpredicting percentile pre-intervention	-1.021	1.853	0.009	-0.104***	-0.145
	(1.270)	(1.433)	(0.024)	(0.033)	(0.195)
	[16.131]	[-12.570]	[0.144]	[0.585]	[9.060]
Overpredicting percentile pre-intervention	-1.666**	-0.116	-0.050***	0.011	-0.192*
	(0.764)	(0.996)	(0.018)	(0.022)	(0.108)
	[19.964]	[12.897]	[0.227]	[0.417]	[8.387]
P-value, treat-by-pre-belief interaction	0.664	0.259	0.047	0.004	0.832
N	2,358	2,358	2,632	2,632	5,715

Notes: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$. Treatment effects estimated from a regression of the outcome on assignment to either treatment, an indicator for whether the student was initially overpredicting their percentile, and a treatment-by-overpredicting interaction, controlling for student academic and demographic characteristics and randomization strata dummies. Initial over vs. underprediciton based on response to item about predicted percentile in the pre-intervention survey. Robust standard errors reported. Outcomes measuring beliefs about percentile and STEM median based on post-intervention survey. Course-taking outcomes based on University of Michigan administrative data.

Table A.18: Estimated Effect of Intervention by Pre-Intervention Error in Prediction of Own Percentile (Continuous)

	Absolute value percentile error	Signed percentile error	Underestimating STEM median	Overestimating STEM median	Number of STEM credits
Treatment (main effect)	-1.163	1.216	-0.013	-0.046**	-0.179
	(0.869)	(0.891)	(0.016)	(0.022)	(0.127)
Treatment*pre-intervention error	-0.036	-0.049	-0.001**	0.002**	-0.002
	(0.033)	(0.032)	(0.001)	(0.001)	(0.004)
N	2,032	2,032	2,223	2,223	3,664

Notes: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$. Treatment effects estimated from a regression of the outcome on assignment to either treatment, a continuous measure of the student's percentile error at the beginning of the semester, and a treatment-by-error interaction, controlling for student academic and demographic characteristics and randomization strata dummies. Initial error is based on response to item about predicted percentile in the pre-intervention survey; a negative error indicates underpredicting, while a positive error indicates overpredicting. Robust standard errors reported. Outcomes measuring beliefs about percentile and STEM median based on post-intervention survey. Course-taking outcomes based on University of Michigan administrative data.

Table A.19: Estimated Effect of Intervention by Student Level

	Absolute value percentile error	Signed percentile error	Underestimating STEM median	Overestimating STEM median	Number of STEM credits
First year or sophomore	-1.574**	0.304	-0.035**	-0.033*	-0.211**
	(0.717)	(0.922)	(0.016)	(0.020)	(0.099)
	[18.767]	[5.956]	[0.208]	[0.471]	[8.580]
Junior or senior	-1.044	2.022	-0.024	0.028	-0.051
	(1.566)	(2.087)	(0.034)	(0.045)	(0.269)
	[20.043]	[8.372]	[0.196]	[0.402]	[8.174]
P-vaue, treat-by-student-level interaction	0.756	0.448	0.776	0.208	0.575
N	2,358	2,358	2,632	2,632	5,715

Notes: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$. Treatment effects estimated from a regression of the outcome on assignment to either treatment, an indicator for whether the student has freshman or sophomore standing, and a treatment-by-level interaction, controlling for student academic and demographic characteristics and randomization strata dummies. Robust standard errors reported. Outcomes measuring beliefs about percentile and STEM median based on post-intervention survey. Student level and course-taking outcomes based on University of Michigan administrative data.

Table A.20: Estimated Effect of Intervention by Pre-Intervention Stated Intended Major

	Absolute value percentile error	Signed percentile error	Underestimating STEM median	Overestimating STEM median	Number of STEM credits
Intended STEM major	-1.860**	0.819	-0.042**	-0.013	-0.248**
	(0.778)	(1.008)	(0.018)	(0.021)	(0.123)
	[19.144]	[5.390]	[0.229]	[0.430]	[9.487]
Intended non-STEM major	0.175	0.675	-0.020	-0.074*	-0.053
	(1.443)	(1.832)	(0.030)	(0.043)	(0.238)
	[17.027]	[7.005]	[0.145]	[0.584]	[4.809]
P-vaue, treat-by-major interaction	0.212	0.945	0.512	0.199	0.466
N	2,165	2,165	2,406	2,406	3,988

Notes: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$. Treatment effects estimated from a regression of the outcome on assignment to either treatment, an indicator for intended STEM major, and a treatment-by-STEM-major interaction, controlling for student academic and demographic characteristics and randomization strata dummies. Intended major based on response to a question about planned major in the pre-intervention survey. Robust standard errors reported. Outcomes measuring beliefs about percentile and STEM median based on post-intervention survey. Course-taking outcomes based on University of Michigan administrative data.

Table A.21: Estimated Effect of Intervention by Whether Student Had Declared a Major at Time of Treatment

	Absolute value percentile error	Signed percentile error	Underestimating STEM median	Overestimating STEM median	Number of STEM credits
Student has declared major	-2.836***	0.070	-0.031	-0.002	-0.014
	(1.058)	(1.370)	(0.025)	(0.027)	(0.141)
	[20.133]	[6.393]	[0.270]	[0.386]	[10.053]
Student undeclared	-0.544	0.956	-0.035**	-0.036	-0.314**
	(0.821)	(1.064)	(0.018)	(0.025)	(0.127)
	[18.187]	[6.339]	[0.163]	[0.508]	[7.290]
P-vaue, treat-by-undecl. interaction	0.084	0.606	0.884	0.345	0.113
N	2,358	2,358	2,632	2,632	5,715

Notes: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$. Treatment effects estimated from a regression of the outcome on assignment to either treatment, an indicator for whether the student was undeclared during the semester of the intervention, and a treatment-by-undeclared interaction, controlling for student academic and demographic characteristics and randomization strata dummies. Robust standard errors reported. Outcomes measuring beliefs about percentile and STEM median based on post-intervention survey. Major status and course-taking outcomes based on University of Michigan administrative data.

Table A.22: Estimated Effect of Intervention by Course Subject

	Absolute value of pctile error	Signed percentile error	Underest. STEM median	Overest. STEM median	Number of STEM credits
Biology	-1.333	0.578	-0.007	-0.075	0.326
	(1.427)	(1.951)	(0.032)	(0.052)	(0.305)
	[16.873]	[5.144]	[0.103]	[0.634]	[7.396]
Chemistry	1.710	-0.749	-0.014	0.017	-0.011
	(2.723)	(3.340)	(0.035)	(0.064)	(0.201)
	[16.963]	[7.630]	[0.033]	[0.817]	[9.534]
Computer Science	-2.295	-2.611	-0.075**	0.028	-0.431*
	(1.697)	(2.227)	(0.038)	(0.043)	(0.250)
	[21.295]	[8.705]	[0.262]	[0.297]	[8.835]
Economics	-1.702	1.152	0.009	-0.071	-0.165
	(2.200)	(2.860)	(0.040)	(0.062)	(0.255)
	[20.041]	[7.694]	[0.102]	[0.648]	[7.007]
Engineering	-5.981***	-0.654	-0.108**	-0.009	0.335
	(1.984)	(2.571)	(0.054)	(0.036)	(0.267)
	[22.992]	[3.938]	[0.561]	[0.108]	[12.763]
Physics	-10.928	2.113	0.098	0.009	-0.082
	(6.774)	(8.431)	(0.108)	(0.143)	(0.367)
	[21.474]	[-4.000]	[0.130]	[0.522]	[12.221]
Statistics	0.446	2.458*	-0.017	-0.027	-0.533***
	(0.998)	(1.278)	(0.022)	(0.032)	(0.197)
	[17.109]	[6.469]	[0.155]	[0.487]	[6.771]
P-vaue, F-test of treat-by-subject interactions	0.060	0.597	0.357	0.738	0.080
N	2,358	2,358	2,632	2,632	5,715

Notes: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$. Treatment effects estimated from a regression of the outcome on assignment to either treatment, course subject, and treatment-by-subject interactions, controlling for student academic and demographic characteristics and randomization strata dummies. Robust standard errors reported. Outcomes measuring beliefs about percentile and STEM median based on post-intervention survey. Course-taking outcomes based on University of Michigan administrative data.

Table A.23: Estimated Effect of Intervention by Gender Composition of Course (Proportion Men, Continuous)

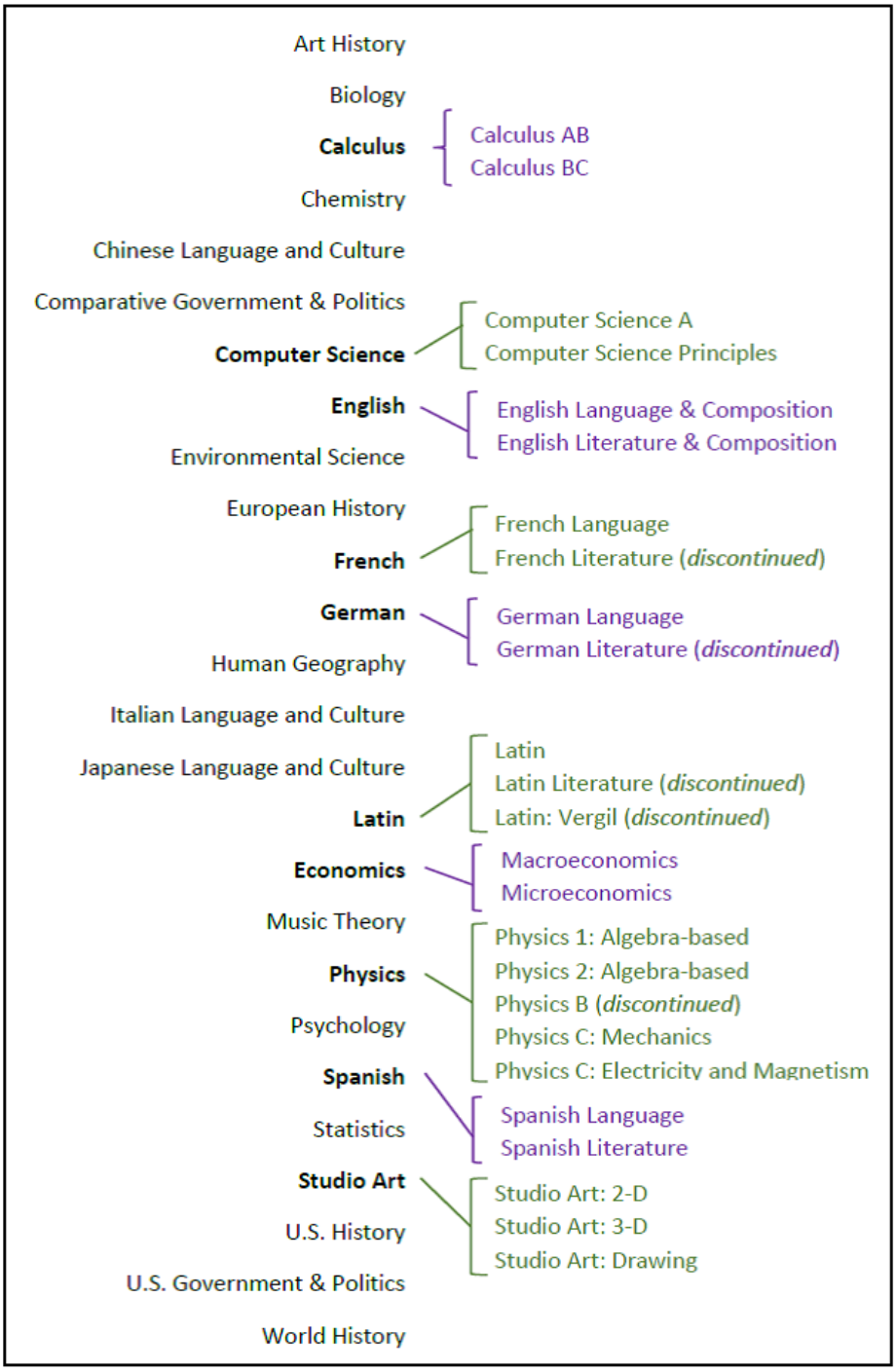
	Absolute value percentile error			Signed percentile error			Underestimating STEM median		
	All	Men	Women	All	Men	Women	All	Men	Women
Treatment-by-proportion-male interaction	-17.451***	-24.341***	-8.222	-9.775	-4.106	-17.300	-0.240*	-0.259	-0.164
	(6.125)	(9.179)	(8.195)	(8.009)	(11.595)	(11.141)	(0.145)	(0.208)	(0.210)
N	2,358	1,166	1,192	2,358	1,166	1,192	2,632	1,291	1,341
	Overestimating STEM median			Number of STEM credits			Took any STEM credits		
	All	Men	Women	All	Men	Women	All	Men	Women
Treatment-by-proportion-male interaction	0.209	0.161	0.148	-0.102	-0.355	0.695	-0.016	-0.069	0.051
	(0.157)	(0.228)	(0.229)	(0.866)	(1.158)	(1.383)	(0.054)	(0.067)	(0.092)
N	2,632	1,291	1,341	5,715	2,993	2,722	5,715	2,993	2,722

Notes: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$. Treatment effects estimated from a regression of the outcome on assignment to either treatment, a continuous measure of the proportion of the course sample that is male (0-1), and a treatment-by-proportion-male interaction, controlling for student academic and demographic characteristics and randomization strata dummies. Effects by gender estimated with a three-way interaction between treatment, a female indicator, and the continuous proportion male. Robust standard errors reported. Outcomes measuring beliefs about percentile and STEM median based on post-intervention survey. Course-taking outcomes based on University of Michigan administrative data.

APPENDIX B

Appendix to The Advanced Placement Program and Educational Inequality

Figure B.1: Coding of AP Courses by Subject



Notes: Subjects in bold have been collapsed from multiple AP subjects, corresponding to the bracketed courses, due to data limitations. This was done because in many cases it was impossible to distinguish, e.g. English Literature and Composition from English Language and Composition (because the school would list the course as "'AP English."

Table B.1: Most Popular AP Courses Offered, Courses Taken, and Exams Taken, by Cohort

	Top 5 AP courses offered (school level)	Top 5 AP courses taken (student level)	Top 5 AP exams taken (student level)
2005	English	English	
	U.S. History	Calculus	
	Calculus	U.S. Government	
	Biology	Biology	
	Chemistry	U.S. History	
2006	English	English	
	Calculus	Calculus	
	U.S. History	U.S. Government	
	Chemistry	Biology	
	Biology	U.S. History	
2007	English	English	English
	Calculus	Calculus	Calculus
	U.S. History	U.S. Government	U.S. Government
	Chemistry	Biology	Biology
	Biology	Psychology	Psychology
2008	English	English	English
	Calculus	Calculus	Calculus
	U.S. History	U.S. Government	U.S. History
	Biology	Biology	Biology
	U.S. Government	Psychology	U.S. Government
2009	English	English	English
	Calculus	Calculus	Calculus
	U.S. History	U.S. Government	U.S. History
	Biology	Biology	Biology
	U.S. Government	Psychology	U.S. Government
2010	English	English	English
	Calculus	Calculus	Calculus
	U.S. History	U.S. Government	U.S. History
	Biology	Biology	U.S. Government
	U.S. Government	Psychology	Biology
2011	English	English	English
	Calculus	Calculus	Calculus
	U.S. History	U.S. Government	U.S. History
	Biology	Psychology	U.S. Government
	U.S. Government	Biology	Biology
2012	English	English	English
	Calculus	Calculus	Calculus
	Biology	Psychology	Psychology
	U.S. History	U.S. Government	U.S. History
	Chemistry	Biology	U.S. Government

Figure B.**2:** Proportion of Schools Offering Any AP Courses, by Cohort

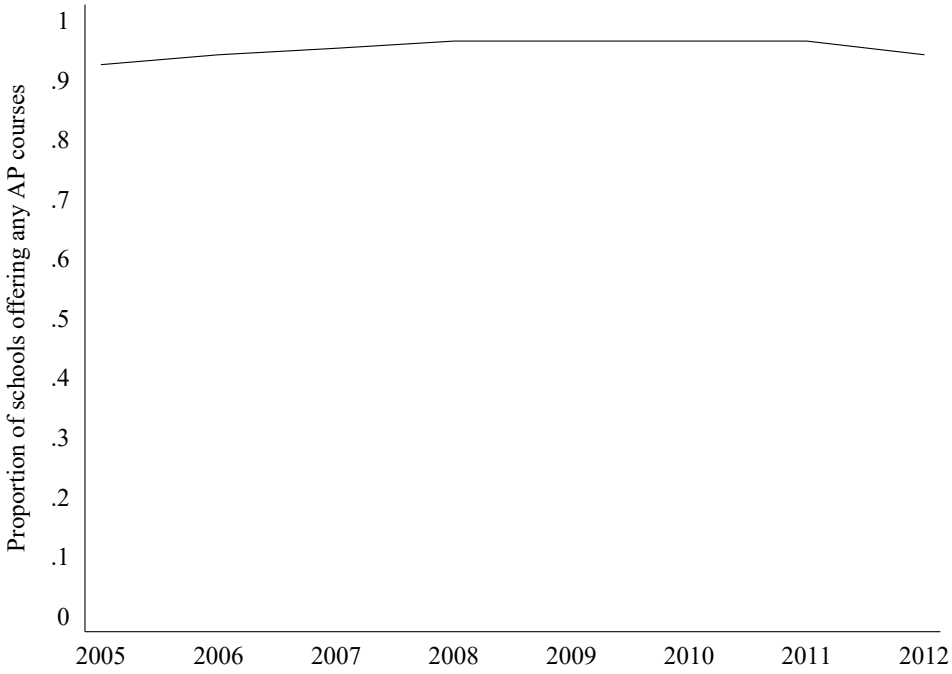


Figure B.**3:** Average Number of AP Courses Offered by School, by Cohort

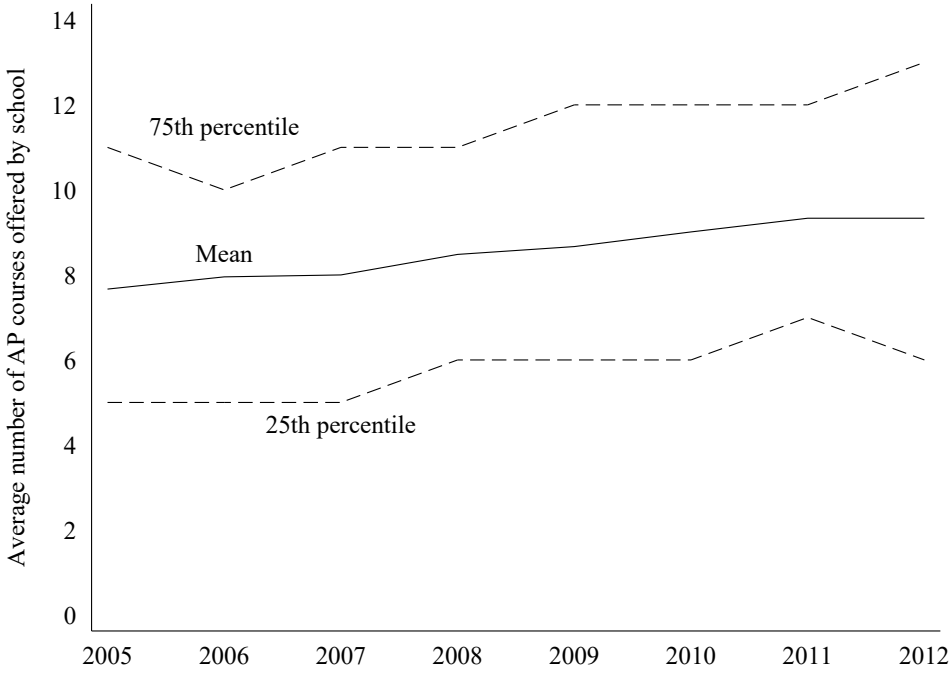


Figure B.**4:** Distribution of Number of AP Courses Offered at a School, by Cohort

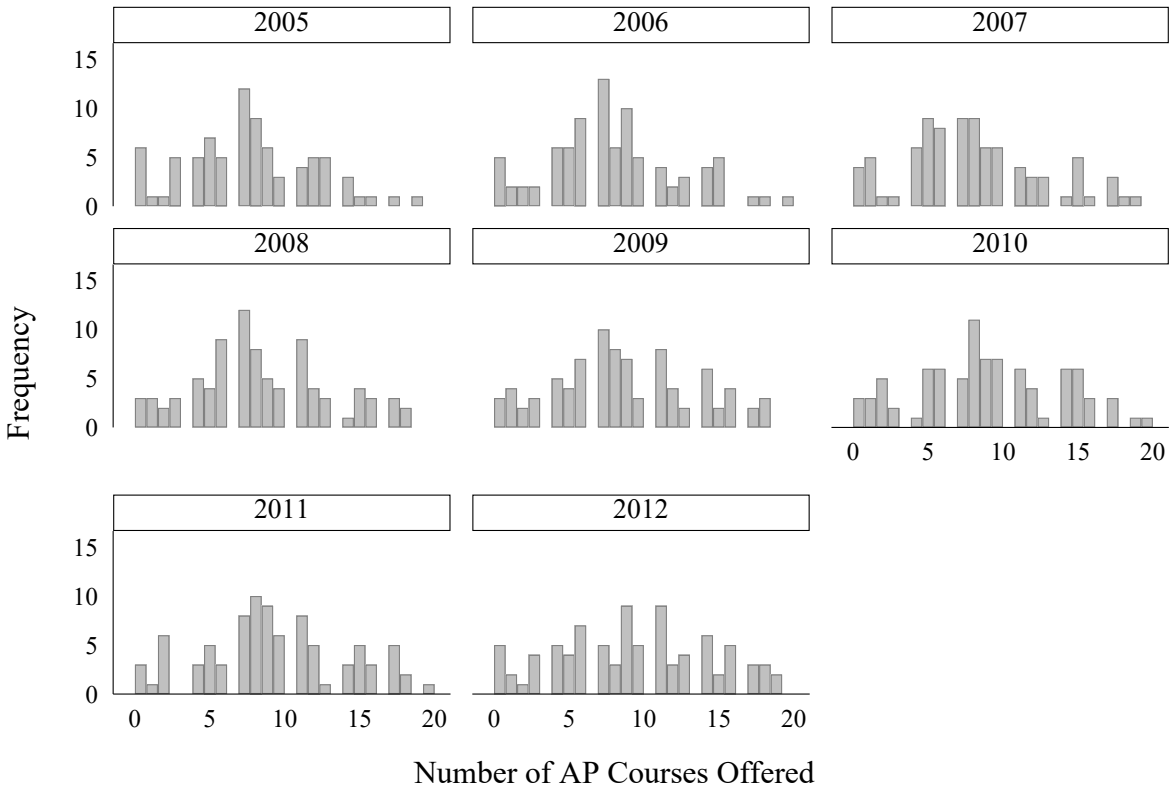


Figure B.**5:** School-by-Cohort Variation in Number of AP Courses Offered

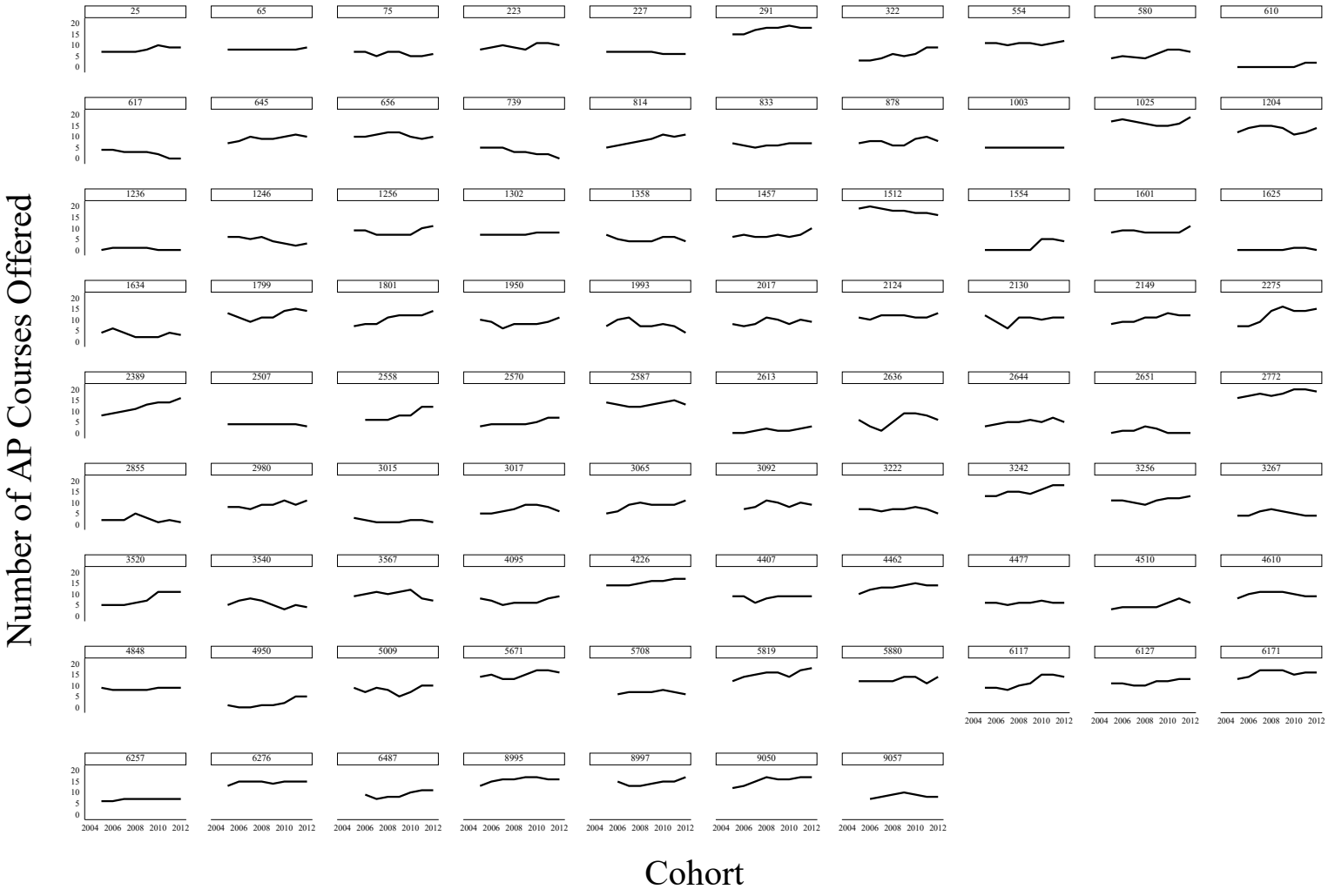


Figure B.**6:** Proportion of Students Taking Any AP Courses, by Cohort and Family Income

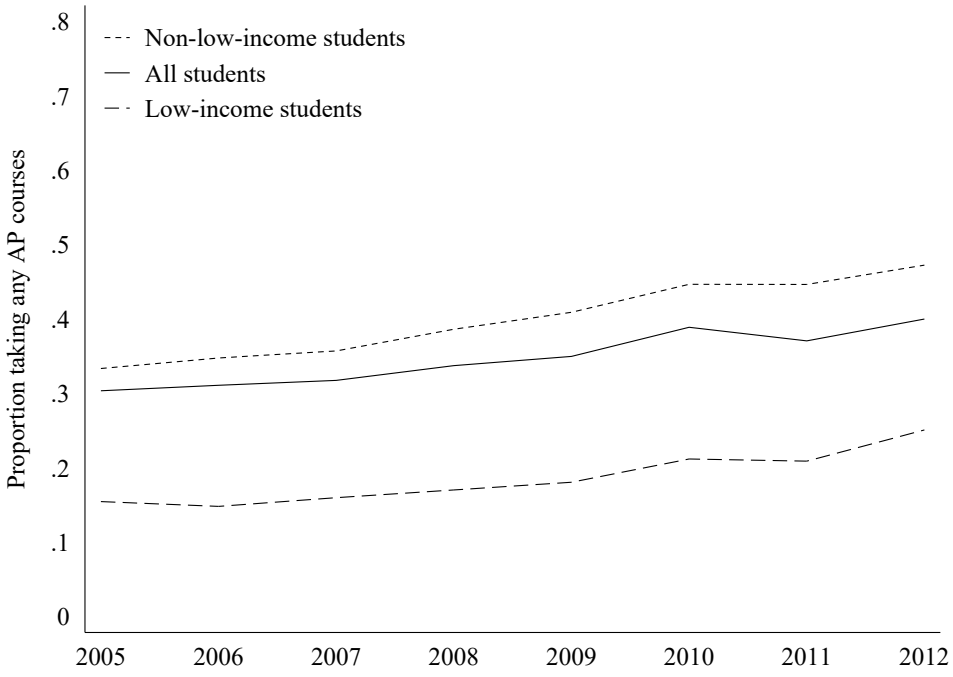


Figure B.**7:** Average Number of AP Courses Taken by Students, Conditional on Taking Any AP, by Cohort and Family Income

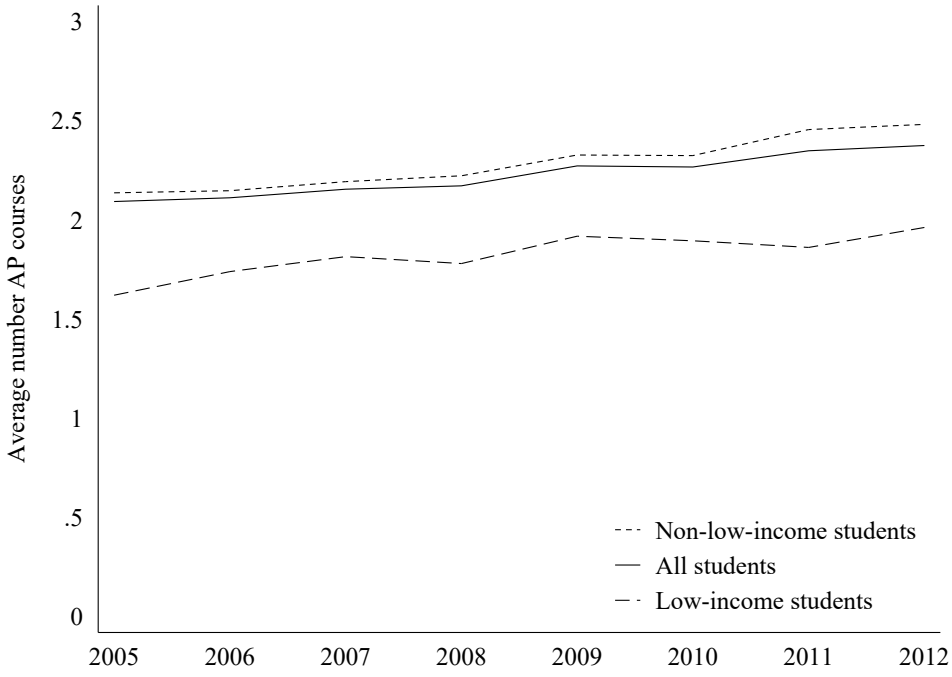


Figure B.**8:** Proportion of Students Taking Any AP Exams, by Cohort and Family Income

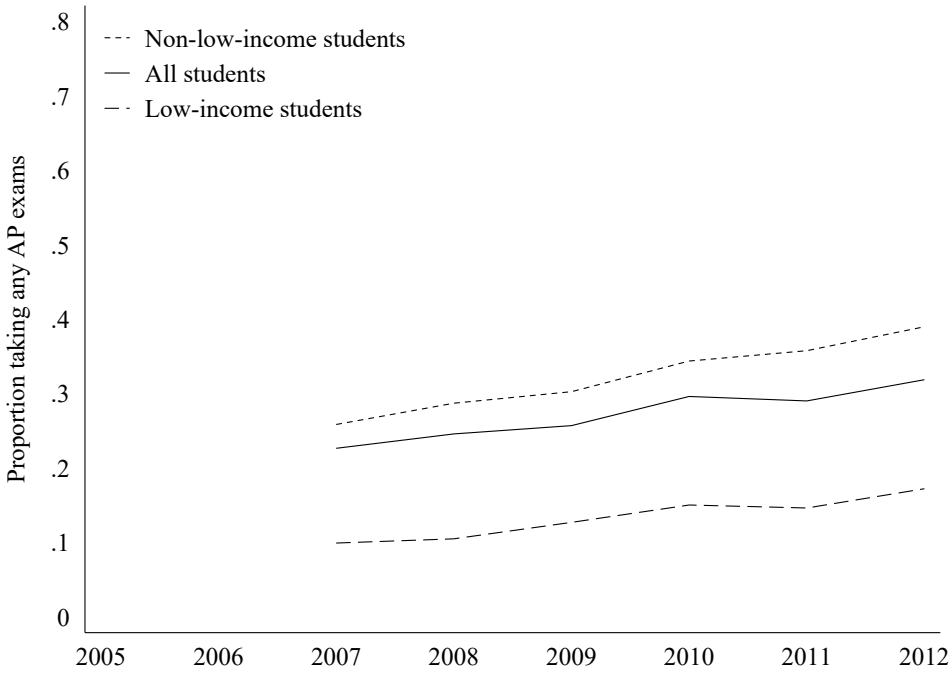


Figure B.**9:** Proportion of Students Taking Any AP Exams, Conditional on Taking Any AP Course, by Cohort and Family Income

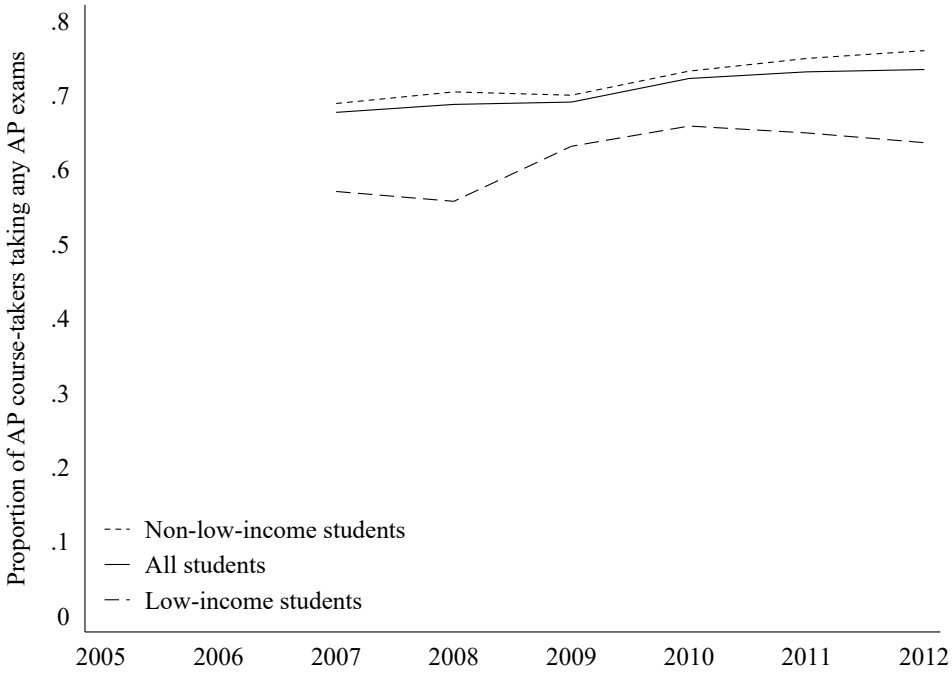
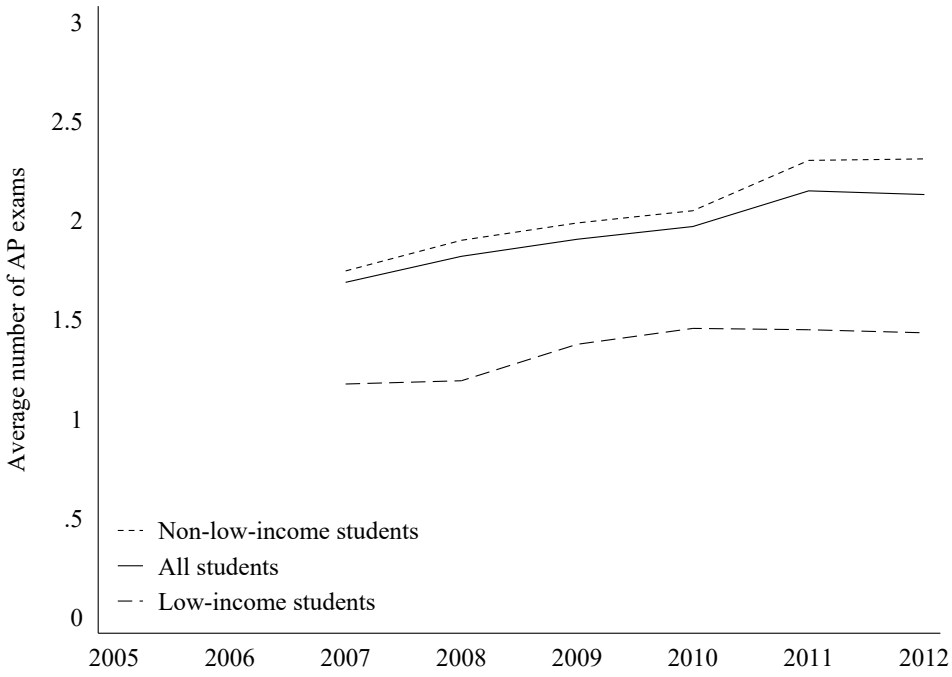


Figure B.**10:** Average Number of AP Exams Taken by Students, Conditional on Taking Any AP Course, by Cohort and Family Income



BIBLIOGRAPHY

BIBLIOGRAPHY

Altonji, J. G. (1993). The demand for and return to education when education outcomes are uncertain. *Journal of Labor Economics*, 11(1, Part 1):48–83.

Altonji, J. G. (1995). The effects of high school curriculum on education and labor market outcomes. *Journal of Human Resources*, pages 409–438.

Altonji, J. G., Arcidiacono, P., and Maurel, A. (2016). The analysis of field choice in college and graduate school: Determinants and wage effects. In *Handbook of the Economics of Education*, volume 5, pages 305–396. Elsevier.

Arcidiacono, P. (2004). Ability sorting and the returns to college major. *Journal of Econometrics*, 121(1-2):343–375.

Arcidiacono, P., Aucejo, E., Maurel, A., and Ransom, T. (2016). College attrition and the dynamics of information revelation. National Bureau of Economic Research Working Paper 22325.

Ashraf, N., Bandiera, O., and Lee, S. S. (2014). Awards unbundled: Evidence from a natural field experiment. *Journal of Economic Behavior & Organization*, 100:44–63.

Astorne-Figari, C. and Speer, J. D. (2019). Are changes of major major changes? The roles of grades, gender, and preferences in college major switching. *Economics of Education Review*, 70:75–93.

Athey, S., Chetty, R., Imbens, G. W., and Kang, H. (2019). The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely. National Bureau of Economic Research Working Paper 26463.

Avery, C., Gurantz, O., Hurwitz, M., and Smith, J. (2018). Shifting college majors in response to Advanced Placement exam scores. *Journal of Human Resources*, 53(4):918–956.

Avilova, T. and Goldin, C. (2020). What can UWE do for economics?. In Lundberg, S., editor, *Women in Economics*. A CEPR Press VoxEU.org book.

Azmat, G., Bagues, M., Cabrales, A., and Iriberri, N. (2019). What you don't know can't hurt you? A natural field experiment on relative performance feedback in higher education. *Management Science*, 65(8):3714–3736.

Azmat, G. and Iriberri, N. (2010). The importance of relative performance feedback information: Evidence from a natural experiment using high school students. *Journal of Public Economics*, 94(7-8):435–452.

Bandiera, O., Larcinese, V., and Rasul, I. (2015). Blissful ignorance? A natural experiment on the effect of feedback on students' performance. *Labour Economics*, 34:13–25.

Bayer, A., Bhanot, S. P., and Lozano, F. (2019). Does simple information provision lead to more diverse classrooms? Evidence from a field experiment on undergraduate economics. *AEA Papers and Proceedings*, 109:110–14.

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

Benjamini, Y., Krieger, A. M., and Yekutieli, D. (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, 93(3):491–507.

Berlin, N. and Dargnies, M.-P. (2016). Gender differences in reactions to feedback and willingness to compete. *Journal of Economic Behavior & Organization*, 130:320–336.

Betts, J. R. (2011). The economics of tracking in education. In *Handbook of the Economics of Education*, volume 3, pages 341–381. Elsevier.

Beyer, S. (1990). Gender differences in the accuracy of self-evaluations of performance. *Journal of Personality and Social Psychology*, 59(5):960.

Beyer, S. and Bowden, E. M. (1997). Gender differences in self-perceptions: Convergent evidence from three measures of accuracy and bias. *Personality and Social Psychology Bulletin*, 23(2):157–172.

Bobba, M. and Frisancho, V. (2019). Perceived ability and school choices. Working paper.

Bordalo, P., Coffman, K., Gennaioli, N., and Shleifer, A. (2016). Stereotypes. *The Quarterly Journal of Economics*, 131(4):1753–1794.

Bordalo, P., Coffman, K., Gennaioli, N., and Shleifer, A. (2019). Beliefs about gender. *American Economic Review*, 109(3):739–73.

Bordón, P., Canals, C., and Mizala, A. (2020). The gender gap in college major choice in Chile. *Economics of Education Review*, 77:102011.

Breda, T. and Napp, C. (2019). Girls' comparative advantage in reading can largely explain the gender gap in math-related fields. *Proceedings of the National Academy of Sciences*, 116(31):15435–15440.

Bruhn, M. and McKenzie, D. (2009). In pursuit of balance: Randomization in practice in development field experiments. *American Economic Journal: Applied Economics*, 1(4):200–232.

Buser, T., Gerhards, L., and Van Der Weele, J. (2018). Responsiveness to feedback as a personal trait. *Journal of Risk and Uncertainty*, 56(2):165–192.

Buser, T., Niederle, M., and Oosterbeek, H. (2014). Gender, competitiveness, and career choices. *The Quarterly Journal of Economics*, 129(3):1409–1447.

Butcher, K. F., McEwan, P. J., and Weerapana, A. (2014). The effects of an anti-grade-inflation policy at Wellesley College. *Journal of Economic Perspectives*, 28(3):189–204.

Calkins, A. (2020). Gender, grades, and college major during the dot-com crash. Working paper.

Ceci, S. J., Ginther, D. K., Kahn, S., and Williams, W. M. (2014). Women in academic science: A changing landscape. *Psychological Science in the Public Interest*, 15(3):75–141.

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

Chizmar, J. F. (2000). A discrete-time hazard analysis of the role of gender in persistence in the economics major. *The Journal of Economic Education*, 31(2):107–118.

Cimpian, J. R., Kim, T. H., and McDermott, Z. T. (2020). Understanding persistent gender gaps in STEM. *Science*, 368(6497):1317–1319.

Coffman, K. B., Collis, M., and Kulkarni, L. (2019). Stereotypes and belief updating. Working paper.

College Board (2003). A brief history of the Advanced Placement program. Retrieved from https://ca01000794.schoolwires.net/cms/lib/CA01000794/Centricity/Domain/295/APUSH/ AP_History_history.pdf.

College Board (2019). Annual AP program participation 1956-2019. Retrieved from https://secure-media.collegeboard.org/digitalServices/pdf/research/2019/2019-Annual-Participation.pdf.

College Board (2020). Program summary report. Retrieved from https://secure-media.collegeboard.org/digitalServices/pdf/research/2020/Program-Summary-Report-2020.pdf.

Conger, D., Kennedy, A. I., Long, M. C., and McGhee, R. (2021). The effect of Advanced Placement science on students skills, confidence, and stress. *Journal of Human Resources*, 56(1):93–124.

Conger, D., Long, M. C., and McGhee Jr, R. (2020). Advanced Placement and initial college enrollment: Evidence from an experiment. Annenberg Institute at Brown University EdWorkingPaper 20-340.

Coutts, A. (2019). Good news and bad news are still news: Experimental evidence on belief updating. *Experimental Economics*, 22(2):369–395.

Darolia, R., Koedel, C., Main, J. B., Ndashimye, J. F., and Yan, J. (2020). High school course access and postsecondary STEM enrollment and attainment. *Educational Evaluation and Policy Analysis*, 42(1):22–45.

Dobrescu, L., Faravelli, M., Megalokonomou, R., and Motta, A. (2019). Rank incentives and social learning: Evidence from a randomized controlled trial. IZA Discussion Paper 12437.

Emerson, T. L., McGoldrick, K., and Mumford, K. J. (2012). Women and the choice to study economics. *The Journal of Economic Education*, 43(4):349–362.

Ertac, S. and Szentes, B. (2011). The effect of information on gender differences in competitiveness: Experimental evidence. Working paper.

Exley, C. L. and Kessler, J. B. (2019). The gender gap in self-promotion. National Bureau of Economic Research Working Paper 26345.

Franco, C. (2019). How does relative performance feedback affect beliefs and academic decisions? Working paper.

Gelbach, J. B. (2016). When do covariates matter? And which ones, and how much? *Journal of Labor Economics*, 34(2):509–543.

Gonzalez, N. (2017). How learning about one's ability affects educational investments: Evidence from the Advanced Placement program. Mathematica Policy Research Working Paper 52.

Goodman, S. (2016). Learning from the test: Raising selective college enrollment by providing information. *Review of Economics and Statistics*, 98(4):671–684.

Goulas, S. and Megalokonomou, R. (2015). Knowing who you are: The effect of feedback information on short and long term outcomes. Working paper.

Gurantz, O. (forthcoming). How college credit in high school impacts postsecondary course-taking: the role of AP exams. *Education Finance and Policy*.

Hsieh, C.-T., Hurst, E., Jones, C. I., and Klenow, P. J. (2019). The allocation of talent and US economic growth. *Econometrica*, 87(5):1439–1474.

Jackson, C. K. (2010). A little now for a lot later: A look at a Texas Advanced Placement Incentive Program. *Journal of Human Resources*, 45(3):591–639.

Jacob, B. A. and Wilder, T. (2010). Educational expectations and attainment. National Bureau of Economic Research Working Paper No. 15683.

Jensen, E. J. and Owen, A. L. (2001). Pedagogy, gender, and interest in economics. *The Journal of Economic Education*, 32(4):323–343.

Kaganovich, M., Taylor, M., and Xiao, R. (2020). Gender differences in persistence in a field of study. Working paper.

Kling, J. R., Liebman, J. B., and Katz, L. F. (2007). Experimental analysis of neighborhood effects. *Econometrica*, 75(1):83–119.

Klopfenstein, K. (2010). Does the Advanced Placement program save taxpayers money? The effect of AP participation on time to college graduation. In Sadler, P. M., Sonnert, G., Tai, R. H., and Klopfenstein, K., editors, *AP: A critical examination of the Advanced Placement program*, pages 189–218. Harvard Education Press.

Klopfenstein, K. and Lively, K. (2012). Dual enrollment in the broader context of college-level high school programs. *New Directions for Higher Education*, 2012(158):59–68.

Klopfenstein, K. and Thomas, M. K. (2010). Advanced placement participation: Evaluating the policies of states and colleges. In Sadler, P. M., Sonnert, G., Tai, R. H., and Klopfenstein, K., editors, *AP: A critical examination of the Advanced Placement program*, pages 167–188. Harvard Education Press.

Kugler, A. D., Tinsley, C. H., and Ukhaneva, O. (2021). Choice of majors: Are women really different from men? *Economics of Education Review*, 81:1–19.

Li, H.-H. (2018). Do mentoring, information, and nudge reduce the gender gap in economics majors? *Economics of Education Review*, 64:165–183.

Li, X. (2019). Duke's introductory economics course set to institute pass/fail grading system. *The Chronicle*. https://www.dukechronicle.com/article/2019/04/duke-econ-economics-101-pass-fail- satisfactory-unsatisfactory-fullenkamp, April 3, 2019.

Lundeberg, M. A., Fox, P. W., and Punćcohaŕ, J. (1994). Highly confident but wrong: Gender differences and similarities in confidence judgments. *Journal of Educational Psychology*, 86(1):114.

Main, J. B. and Ost, B. (2014). The impact of letter grades on student effort, course selection, and major choice: A regression-discontinuity analysis. *The Journal of Economic Education*, 45(1):1–10.

Malkus, N. (2016). The AP peak: Public schools offering Advanced Placement, 2000-12. AEI Papers & Studies. Retrieved from https://www.aei.org/research-products/report/the-ap-peak-public-schools-offering-advanced-placement-2000-12/.

Marshman, E. M., Kalender, Z. Y., Nokes-Malach, T., Schunn, C., and Singh, C. (2018). Female students with A's have similar physics self-efficacy as male students with C's in introductory courses: A cause for alarm? *Physical Review Physics Education Research*, 14(2):020123.

Mobius, M. M., Niederle, M., Niehaus, P., and Rosenblat, T. S. (2014). Managing self-confidence. Working paper.

Niederle, M. and Vesterlund, L. (2007). Do women shy away from competition? Do men compete too much? *The Quarterly Journal of Economics*, 122(3):1067–1101.

Niederle, M. and Vesterlund, L. (2011). Gender and competition. *Annual Review of Economics*, 3(1):601–630.

Olson, S. and Riordan, D. G. (2012). Engage to excel: Producing one million additional college graduates with degrees in science, technology, engineering, and mathematics. Report to the President, Executive Office of the President.

Ost, B. (2010). The role of peers and grades in determining major persistence in the sciences. *Economics of Education Review*, 29(6):923–934.

Owen, A. L. (2010). Grades, gender, and encouragement: A regression discontinuity analysis. *The Journal of Economic Education*, 41(3):217–234.

Porter, C. and Serra, D. (2019). Gender differences in the choice of major: The importance of female role models. *American Economic Journal: Applied Economics*, 12(3):226–254.

Rask, K. and Tiefenthaler, J. (2008). The role of grade sensitivity in explaining the gender imbalance in undergraduate economics. *Economics of Education Review*, 27(6):676–687.

Rask, K. N. and Bailey, E. M. (2002). Are faculty role models? Evidence from major choice in an undergraduate institution. *The Journal of Economic Education*, 33(2):99–124.

Rodriguez, A. and McGuire, K. M. (2019). More classes, more access? Understanding the effects of course offerings on Black-White gaps in Advanced Placement course-taking. *The Review of Higher Education*, 42(2):641–679.

Rodriguez, A., McKillip, M. E., and Niu, S. X. (2013). The earlier the better? Taking the AP in 10th grade. College Board Research Report No. 2012-10.

Roy, A. D. (1951). Some thoughts on the distribution of earnings. *Oxford Economic Papers*, 3(2):135–146.

Schneider, J. (2009). Privilege, equity, and the Advanced Placement program: Tug of war. *Journal of Curriculum Studies*, 41(6):813–831.

Smith, J., Hurwitz, M., and Avery, C. (2017). Giving college credit where it is due: Advanced Placement exam scores and college outcomes. *Journal of Labor Economics*, 35(1):67–147.

Solorzano, D. G. and Ornelas, A. (2002). A critical race analysis of Advanced Placement classes: A case of educational inequality. *Journal of Latinos and Education*, 1(4):215–229.

Stinebrickner, R. and Stinebrickner, T. R. (2014). A major in science? Initial beliefs and final outcomes for college major and dropout. *Review of Economic Studies*, 81(1):426–472.

Stinebrickner, T. and Stinebrickner, R. (2012). Learning about academic ability and the college dropout decision. *Journal of Labor Economics*, 30(4):707–748.

Stinebrickner, T. R. and Stinebrickner, R. (2011). Math or science? Using longitudinal expectations data to examine the process of choosing a college major. National Bureau of Economic Research Working Paper16869.

Theokas, C. and Saaris, R. (2013). Finding America's missing AP and IB students. The Education Trust Shattering Expectations Series. Retrieved from https://edtrust.org/wp-content/uploads/2013/10/Missing_Students.pdf.

Tran, A. and Zeckhauser, R. (2012). Rank as an inherent incentive: Evidence from a field experiment. *Journal of Public Economics*, 96(9-10):645–650.

U.S. Department of Education (2016). U.S. Education Department awards 41 states and the District of Columbia $28.4 million in grants to help students from low-income families take Advanced Placement tests. Press release. Retrieved from https://www.ed.gov/news/press-releases/us-education-department-awards-41-states-and-district-columbia-284-million-grants-help-students-low-income-families-take-advanced-placement-tests.

Vincent-Ruz, P., Binning, K., Schunn, C. D., and Grabowski, J. (2018). The effect of math SAT on women's chemistry competency beliefs. *Chemistry Education Research and Practice*, 19(1):342–351.

Webber, D. A. (2019). Projected lifetime earnings by major. Technical report.

Westfall, P. H. and Young, S. S. (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment*, volume 279. John Wiley & Sons.

Wozniak, D., Harbaugh, W. T., and Mayr, U. (2014). The menstrual cycle and performance feedback alter gender differences in competitive choices. *Journal of Labor Economics*, 32(1):161–198.

Xue, Y. and Larson, R. C. (2015). STEM crisis or STEM surplus? Yes and yes. *Monthly Labor Review*.

Zafar, B. (2011). How do college students form expectations? *Journal of Labor Economics*, 29(2):301–348.

Zafar, B. (2013). College major choice and the gender gap. *Journal of Human Resources*, 48(3):545–595.