College Field Specialization and Beliefs about Relative Performance

An Experimental Intervention to Understand Gender Gaps in STEM

[Click here for latest version]

Stephanie Owen *

October 29, 2020

Abstract

Beliefs about relative academic performance may be a determinant of field specialization and may explain some of the gender gap in college STEM enrollment, but there is little causal evidence isolating this mechanism. To test whether beliefs about relative performance are malleable and salient enough to change behavior, I run a randomized controlled trial with 5,700 undergraduate students across seven STEM disciplines. Treated students receive information about their performance relative to their classmates and to STEM majors. 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.

*University of Michigan Department of Economics and Ford School of Public Policy. Contact: srowen@umich.edu. I thank my advisers for their invaluable support and guidance: Sue Dynarski, Sara Heller, Kevin Stange, and Charlie Brown. This work benefited from numerous conversations with friends and peers: Mattan Alalouf, Elizabeth Burland, Jamie Fogel, Matthew Gross, Max Gross, Thomas Helgerman, Shawn Martin, Dhiren Patki, Shwetha Raghuraman, Michael Ricks, Anna Shapiro, Andrew Simon, Sabrina Solanki, Ellen Stuart, and Brittany Vasquez. Peter Blair, Sarah Cohodes, Ashley Craig, and Basit Zafar provided helpful comments on an early draft. I am grateful for feedback from seminar and conference participants at the University of Michigan (especially the Causal Inference in Education Research Seminar), the Association for Education Policy and Finance, and the Health, History, Demography, and Development Research Day. I gratefully acknowledge financial support from the U.S. Department of Education's Institute of Education Sciences through PR/Award R305B150012#. This project would not have been possible without the ECoach research team within the University of Michigan's Center for Academic Innovation, especially Holly Derry, Ben Hayward, Caitlin Hayward, Tim McKay, and Kyle Schulz.

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 debias 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.

[1] Throughout the paper, references to STEM include economics.

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 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, suggest 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 Stinebricker 2011, 2012, and 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 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.

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). By continuing to follow the students in my study, 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 2, describe my experimental design in Section 3, and lay out my empirical methods in Section 4. In section 5 I document baseline gender differences in beliefs about relative performance. Section 6 includes the experimental results of my intervention. Section 7 contextualizes the results and Section 8 concludes.

[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.

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 and 17,000 graduate students. I focus on 5,715 undergraduate students enrolled in seven large introductory STEM courses in 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 applicants, such as international and transfer students, do not submit this information.

The second source is a set of surveys that I administered to all students in the sample

[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).

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 for the post-intervention survey.

3 Experimental Design

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

[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.

[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.

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!" 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

[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.

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.

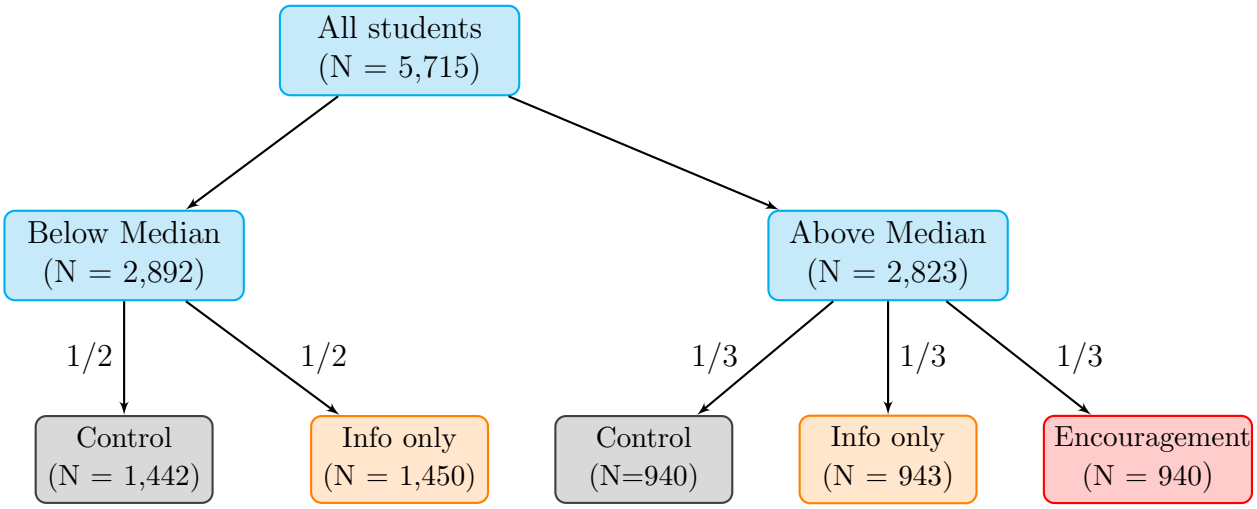
3.2 Treatment assignment

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 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

[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.

Figure 1: Experimental Design



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 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).

3.3 Sample characteristics and balance

Table 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 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 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 t-test by 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

[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.

[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.

Table 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.

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, 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 3.2).

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 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 than others 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.

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 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

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

than the reference group of American Indian or multiple race students. Finally, 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 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.

4 Empirical Method

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 my 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.

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:

$$Y_i = \beta_0 + \beta_1 Treat_i + \boldsymbol{\gamma X_i'} + \delta_s + \epsilon_i \tag{1}$$

where $Treat_i$ indicates assignment to the either treatment, X_i is a vector of pre-treatment covariates (everything listed in Table 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:

$$Y_i = \beta_0 + \beta_1 Female_i + \beta_2 Treat_i + \beta_3 Female_i \cdot Treat_i + \boldsymbol{\gamma X_i'} + \delta_s + \epsilon_i \tag{2}$$

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

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

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:

$$Y_i = \beta_0 + \beta_1 Info_i + \beta_2 Encourage_i + \boldsymbol{\gamma X_i'} + \delta_s + \epsilon_i \tag{3}$$

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 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 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 (see 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.

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.

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

[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.

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 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 (under- predicting by an average of 5.8 percentiles compared to less than a percentile for high-performing men).[18]

Panel (a) of Figure 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.

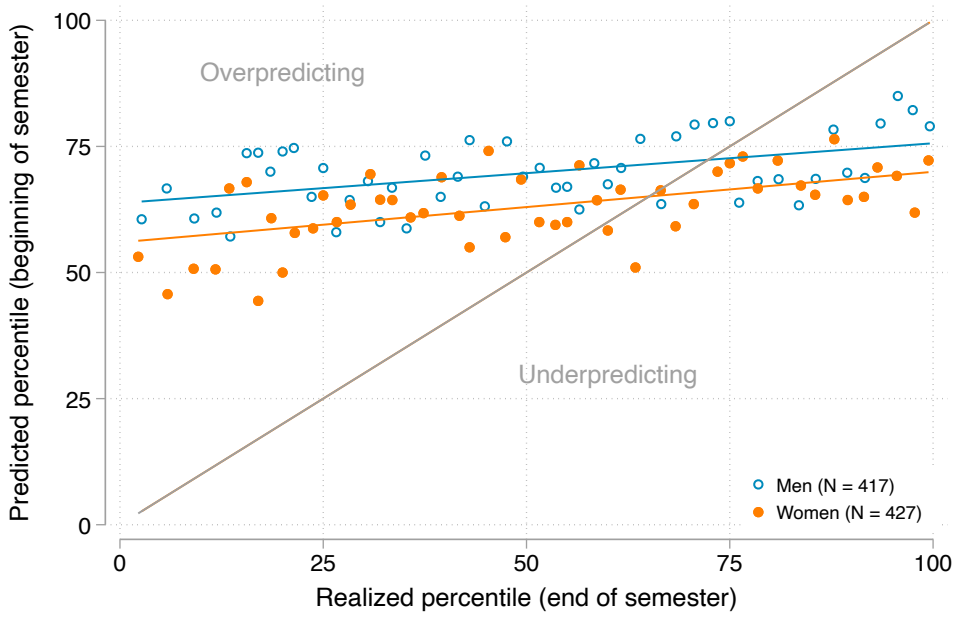
Even absent intervention, we would expect students to update their beliefs over the course of the semester as they learn about their performance through exams, assignments, and other feedback. At the end of the semester (right before final exams), control students' predictions are more accurate than 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.

[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.

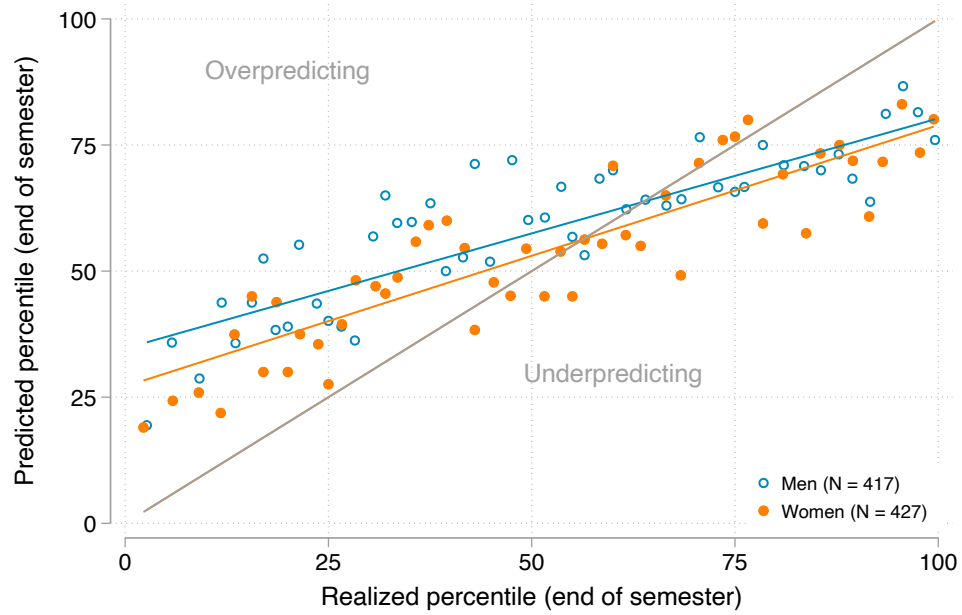
Compared to an absolute value error of 28 percentiles at the beginning of the semester, the average control student's absolute error at the end of the semester is 19.2. The fact that the change in the signed error is similar to the change in the absolute value of the error suggests that it is primarily 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 ($p<0.1$).

These changes are reflected in Panel (b) of Figure 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.

Figure 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.

5.2 Student beliefs about other STEM majors

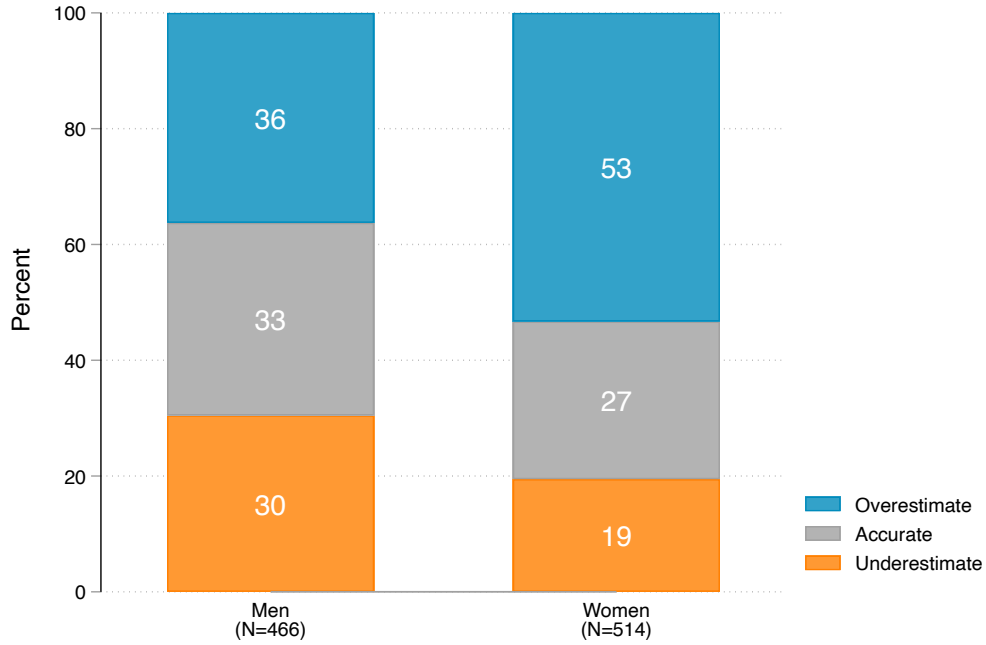
My second measure of beliefs about relative performance focuses on what students believe about STEM majors. I ask students what they think the median grade in their course is among students who go on to major in a STEM field; I can then compare their answers to the true median.[19] This measure captures 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 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.

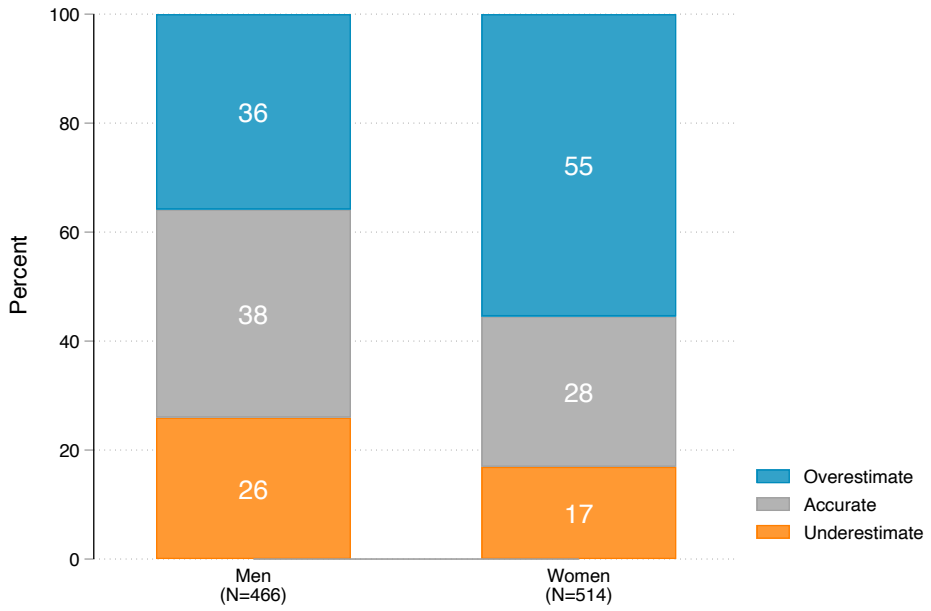
Control students' beliefs about the course median for STEM majors change little over the semester (Panel (b) of Figure 3). This is unsurprising; though they learn about their own performance and, to a lesser extent, that of their peers, they receive no direct information about STEM majors' grades in 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.

Figure 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.

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.

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]

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, 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 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

[20]The Gelbach decomposition is conceptually similar to a Kitagawa-Oaxaca-Blinder decomposition, and in fact is equivalent once interactions between the covariates and gender are included.

Table 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.

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.

6 Experimental Results

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 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 controls means in the treatment effects tables differ from the values reported in Section 5.1.) I test for effects on performance directly in Section 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

[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 surpising given that the intervention told students their mid-semester percentile; they updated their beliefs in the direction of the signal they received.

indicator variables for whether a student is over- or underestimating the course median for students who go on to major in STEM.

Table 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. 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.

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

Both changes suggest that men are becoming less overconfident relative to women.

In Table 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 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).

Table 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 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.

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.)

6.2 Effect of intervention on STEM persistence

As outlined in my pre-analysis plan, my primary outcomes are three measures of persistence in STEM, operationalized as a short-term, medium-term, and long-term measure. In this draft, I focus on the short-term primary outcome, enrollment in STEM courses, because only one semester of data is available since the intervention. I will eventually be able to measure 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.

The primary outcome I present is enrollment in STEM courses, which I operationalize as the number of credits a student attempted in 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, which the University of Michigan assigns to all courses.[23] The following subjects (CIP codes) are considered STEM: natural resources and conservation (03), computer and information sciences (11) engineering (14), biological

[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).

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]

Table 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 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 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,

[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.

Table 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.

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 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]

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. Since only one semester has passed since the experiment, 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 aslo 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

[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 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.

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 women are small, negative, and not statistically signficantly 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

[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.

persistence and major choice are likely to be small.

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 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

[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.

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 6. The effects of the intervention on this success 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. 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.

Table 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.

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 Tables 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 academic fields with similar characteristics, there is considerable variation in key factors such as the proportion of women and the mathematical intensity across fields. The seven courses in my study vary in these ways as well as in course content, grading structure, and more. 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), 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 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 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

[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 percent), and Biology (35 percent).

interested in STEM) change their behavior the most.

7 Discussion

This work fits within 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 (see, e.g., 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 (see, e.g., Mobius et al. 2014 or 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 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 men are incorporating signals from other sources like standardized test scores and STEM courses they took previously, and they 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 (Bordalo et al. 2016), due to people using mental shortcuts to make predictions about themselves or others. 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.

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 STEM 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.

References

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

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

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

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

Athey, S., R. Chetty, G. W. Imbens, and H. Kang (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.

Azmat, G., M. Bagues, A. Cabrales, and N. Iriberri (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 N. Iriberri (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., V. Larcinese, and I. Rasul (2015). Blissful ignorance? A natural experiment on the effect of feedback on students' performance. *Labour Economics 34*, 13–25.

Bayer, A., S. P. Bhanot, and F. Lozano (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.

Beede, D. N., T. A. Julian, D. Langdon, G. McKittrick, B. Khan, and M. E. Doms (2011). Women in STEM: A gender gap to innovation. *Economics and Statistics Administration Issue Brief* (04-11).

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

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

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 E. M. Bowden (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 V. Frisancho (2019). Perceived ability and school choices. Working paper.

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

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

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

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

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

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

Buser, T., N. Peter, and S. C. Wolter (2017). Gender, competitiveness, and study choices in high school: Evidence from Switzerland. *American Economic Review 107*(5), 125–30.

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

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

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

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

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

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

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

Exley, C. L. and J. B. Kessler (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.

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

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

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

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

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

Manski, C. F. (2004). Measuring expectations. *Econometrica 72*(5), 1329–1376.

Marshman, E. M., Z. Y. Kalender, T. Nokes-Malach, C. Schunn, and C. Singh (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., M. Niederle, P. Niehaus, and T. S. Rosenblat (2014). Managing self-confidence. Working paper.

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

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

Olson, S. and D. G. Riordan (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.

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

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

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

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

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

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

Vincent-Ruz, P., K. Binning, C. D. Schunn, and J. Grabowski (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 S. S. Young (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment*, Volume 279. John Wiley & Sons.

Wiswall, M. and B. Zafar (2015). Determinants of college major choice: Identification using an information experiment. *The Review of Economic Studies 82*(2), 791–824.

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

Xue, Y. and R. C. Larson (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.

Appendix Tables and Figures

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	Treatment	p-value	Control	Treatment	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 miissing	-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	Took survey coef.	Characteristic	Took survey 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 miissing	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 3 and 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 3 and 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 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 5. Robust standard errors reported. Course-taking outcomes based on University of Michigan administrative data.

Table A.12: Estimatd 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 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 3.

Table A.15: Estimated Effect of Intevention 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 Prediciton 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-vaue, 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 Prediciton 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.

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

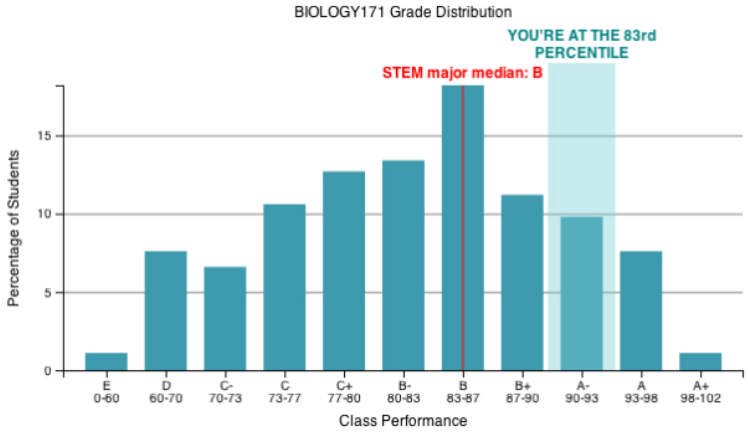
Your Bio 171 grade | And your major

[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

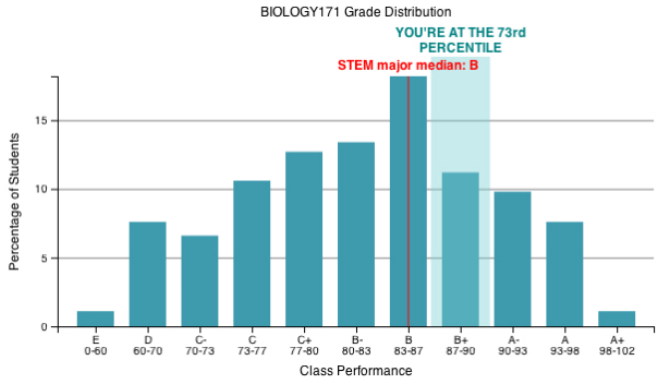
[redacted]

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:

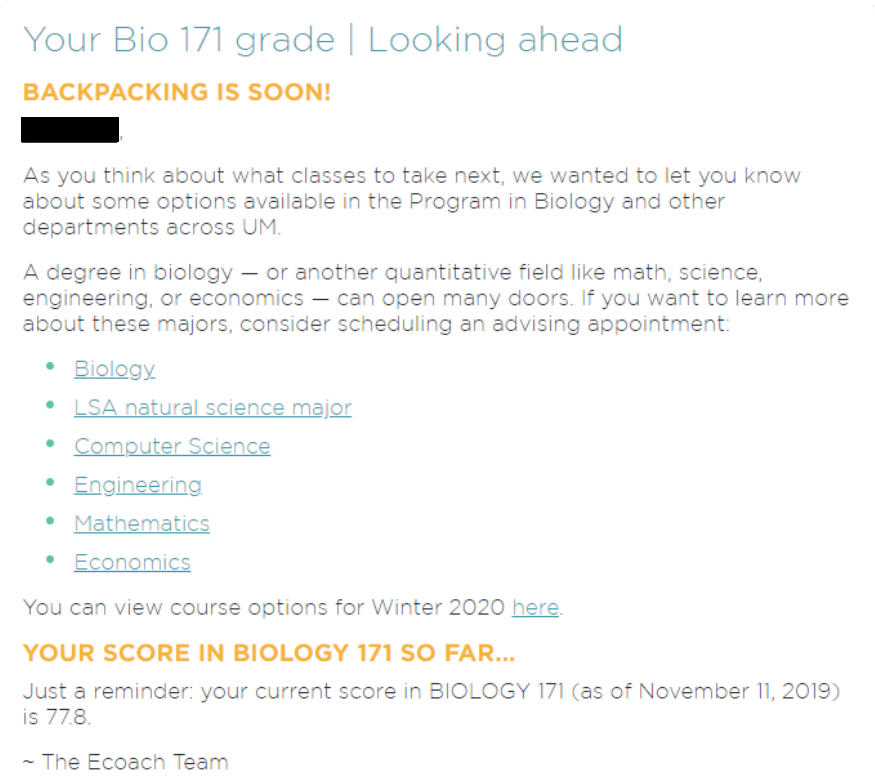
- Biology
- LSA natural science major
- Computer Science
- Engineering
- Mathematics
- Economics

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