ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 22-604

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

Stephanie Owen
Colby College

Beliefs about relative academic performance may shape field specialization and explain gender gaps in STEM enrollment, but little causal evidence exists. To test whether these beliefs are malleable and salient enough to change behavior, I run a randomized controlled trial with 5,700 undergraduates across seven introductory STEM courses. Providing relative performance information shrinks gender gaps in biased beliefs substantially and closes ten percent of the gender gap in subsequent STEM course-taking. The gap closes due to men taking fewer STEM credits; women's behavior is unchanged, implying that male overconfidence rather than female underconfidence contributes to gaps in specialization. Beliefs matter, but may not be a useful target for facilitating female STEM participation.

VERSION: July 2022

Suggested citation: Owen, Stephanie. (2022). College Field Specialization and Beliefs about Relative Performance: An Experimental Intervention to Understand Gender Gaps in STEM. (EdWorkingPaper: 22-604). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/webn-jz41

College Field Specialization and Beliefs about Relative Performance

An Experimental Intervention to Understand Gender Gaps in STEM

Stephanie Owen*

Colby College

July 7, 2022

Abstract

Beliefs about relative academic performance may shape field specialization and explain gender gaps in STEM enrollment, but little causal evidence exists. To test whether these beliefs are malleable and salient enough to change behavior, I run a randomized controlled trial with 5,700 undergraduates across seven introductory STEM courses. Providing relative performance information shrinks gender gaps in biased beliefs substantially and closes ten percent of the gender gap in subsequent STEM course-taking. The gap closes due to men taking fewer STEM credits; women's behavior is unchanged, implying that male overconfidence rather than female underconfidence contributes to gaps in specialization. Beliefs matter, but may not be a useful target for facilitating female STEM participation.

Keywords— STEM, beliefs, gender, college major

*Contact: sowen@colby.edu. I thank Sue Dynarski, Sara Heller, Kevin Stange, and Charlie Brown for their invaluable support and guidance. This work benefited from numerous conversations with colleagues at the University of Michigan. Peter Blair, Sarah Cohodes, Ashley Craig, Amanda Griffith, and Basit Zafar provided helpful comments on early drafts. I am grateful for feedback from seminar and conference participants at the University of Michigan, the University of Chicago, the Association for Education Policy and Finance, the Association for Public Policy Analysis and Management, and the Liberal Arts College Labor and Public Conference. 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. 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.

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 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.[1]

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 specialization contribute to the gender pay gap. Median lifetime earnings for economics or computer engineering majors—fields where men are overrepresented—are roughly 40 percent higher than those for English or psychology majors—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 gender 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; Exley and Kessler, 2019), while economic theory predicts that beliefs about field-specific ability are a determinant of field specialization (Altonji et al., 2016; Arcidiacono, 2004). Research from the lab and the field has shown that information provision can de-bias beliefs and change behavior in a variety of settings (Wozniak et al., 2014; Bobba and Frisancho, 2019; Franco, 2019; Gonzalez, 2017). Several recent field experiments have shown that it is possible to change the academic decisions of college students with light-touch interventions, though cannot disentangle the mechanisms responsible or the reasons for gender differences (Li, 2018; Porter and Serra, 2019; Bayer et al., 2019).[2] Together, these prior strands of work suggest that beliefs

[1] Author's calculations using 2017 IPEDS data.

[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

about performance may be malleable and salient enough to affect the college field specialization choices of underrepresented groups, but causal evidence on this mechanism has thus far been limited.

This paper provides 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.[3] 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.

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 four percentiles more than women do; this is due more to overconfidence of low-performing men than underconfidence of women.

---

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.

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

I also find striking and persistent gender differences in students' accuracy in identifying the median course grade for students who go on to major in STEM. Men are about ten percentage points more likely to think the median course grade for students who go on to major in STEM is lower than it actually is, while women are about 20 percentage points more likely to think it is higher than it is. The patterns in this second type of belief, which no other study has measured, imply male overconfidence and female underconfidence about their performance relative to others.

Providing information on true relative performance closes the gender gap in STEM persistence modestly. The intervention 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; women do not change their course-taking at all.

Survey evidence suggests that the informational intervention caused students to revise their 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. The results by gender are broadly consistent with 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.

Further disaggregation of results by pre-intervention beliefs and performance reveals a somewhat more nuanced story, with both low- and high-performing men decreasing their STEM course-taking. Furthermore, patterns by pre-intervention beliefs are not fully consistent with a model of belief updating. Due the lower statistical power, I caution against overinterpreting these less well-powered analyses. Furthermore, unlike similar studies in the lab, my measures of beliefs are taken several months apart so that pre-intervention beliefs do not perfectly capture beliefs at the time of treatment. To the extent that students update their beliefs between initially being surveyed and being treated, heterogeneity at this level will be muted.

Additional heterogeneity analysis indicates that students we might expect to be on the margin of switching —those already interested in STEM and earlier in their college experience—change their behavior the most. The intervention does not affect students' 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. I detect no differences by treatment arm on course-taking behavior. For this reason, the majority of the results I present combine the two treatment arms and reflect a general effect of information provision.

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 rather than female underconfidence. 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, and 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 male peers). Though prior work put forward beliefs as a promising mechanism, my results imply that beliefs may not be a useful target for increasing female STEM participation.

The paper proceeds as follows. I introduce the setting and data in Section 2, describe the experiment in Section 3, and lay out empirical methods in Section 4. I present my results in Section 5. Section 6 contextualizes the results and Section 7 concludes.

2 Setting, Data, and Sample

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

[4] A second round of the study, planned for the spring semester (referred to as the winter term at the University of Michigan) of 2020, was canceled due to the COVID-19 pandemic.

[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).

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 (University of Michigan Office of Enrollment Management, 2021). 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, and official grades. Because these are administrative data, they contain full information on academic outcomes for all students. Some students are missing information on pre-college characteristics such as high school GPA and parental education, which is collected as part of the application process. This is because some information, such as parental education, is self-reported, and some applicants, such as international and transfer students, do not submit certain information.

The second source is a set of surveys that I administered to all students in the sample at two points in time: one survey before the intervention and one after the intervention (University of Michigan Center for Academic Innovation, 2019). 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.

[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.

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 were 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.

The first treatment arm, the information-only intervention, provides students with information about their performance relative to their classmates and to STEM majors. The message includes a histogram showing the current distribution of grades in the course. The student's own grade is highlighted and their percentile is labeled (e.g., "You're at the 75th percentile"). The graph also includes a call-out informing students about the typical grade in the course for a STEM major (e.g., "STEM major median: B+"). All of the key information in the chart—the student's score and percentile and the median for STEM majors—is repeated later in the message. The second part of the message gives further context about grades in the course, listing the course median for all students, students who go on to major in the field associated with the course,[9] and (again) students who go on to major in STEM. 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, 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.[10] Appendix Figure A.2 shows an example of an information-plus-encouragement 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.

[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.

[10]If the student indicated on the pre-intervention survey that they intended to major in a STEM field, they were encouraged to stay in their major; if they did not (or did not answer) they were urged to consider a STEM field.

In designing a second treatment arm, I wanted to test whether the framing of the information affected how students incorporated it. The findings of Li (2018), an experimental intervention that bundled relative performance information with encouragement and information about the field of economics, suggest that the encouragement aspect may be important for high-performing women in particular but cannot disentangle the various components.[11]

Students already know (or can easily see) 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 received messages informing them of their current score, but 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.

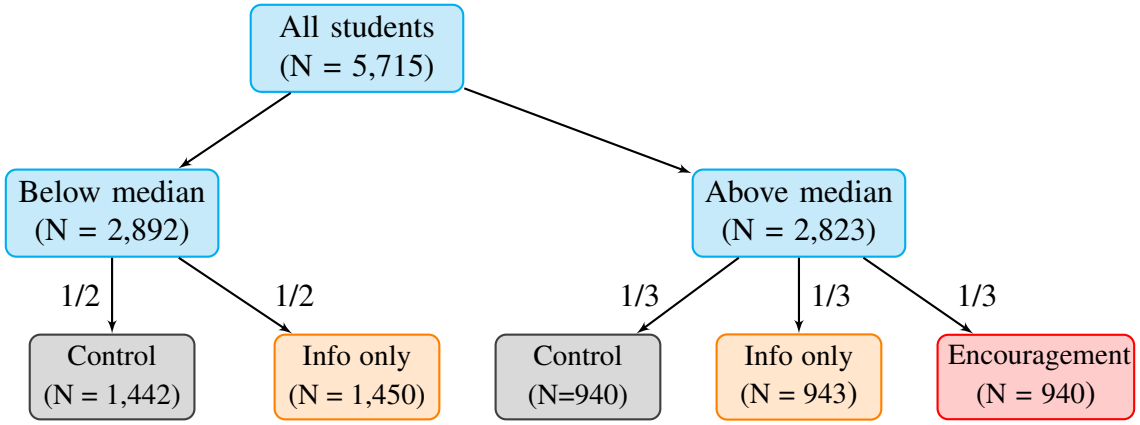
3.2 Treatment Assignment

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.[12] 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 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

[11] 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. Because these three elements were bundled, 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.

[12] 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



Notes: Median is in reference to the course-specific distribution (e.g., the median for STATS 250). “Info only" refers to the information-only treatment; “Encouragement" refers to the information-plus-encouragement treatment arm.

re-randomization and its implications for inference in my analysis by using randomization-based inference in addition to standard model-based inference; see Appendix B). This randomization 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.[13]

3.3 Sample Characteristics and Balance

Table 1 summarizes demographic and academic characteristics by treatment status. This table also tests for balance on pre-treatment characteristics between control students and treated students.[14]

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 have first year or sophomore standing (42 and 40 percent, respectively).[15] The average UM student and the

[13]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).

[14]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.

[15]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.

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.[16]

I test for balance on each pre-treatment characteristic, as well as for the proportion of students missing information on each characteristic, with a regression of the characteristic on treatment status, controlling for strata. I find one significant difference out of 36 tests, fewer than would be expected by chance. Treated students have an average ACT reading subscore that is 0.1 points lower on the 36-point scale, which is substantively small. I also test for whether the characteristics jointly predict treatment status, again controlling for strata; the p-value from this F-test is 0.836.

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. The full breakdown of the sample by course and gender is available as Appendix Table A.3.

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

[16]Engineering is its own college and prospective engineers are admitted directly into the program as incoming first years, meaning engineering majors enter UM already declared. 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	0.318	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	0.422	
Asian	0.254	0.289	0.156	
Black	0.038	0.025	0.212	
Declared major (omitted: other)				
Undeclared	0.560	0.559	0.606	5,715
Engineering	0.232	0.236	0.484	
Math, science, or economics	0.095	0.094	0.657	
Academic and demographic characteristics				
In-state	0.524	0.520	0.362	5,715
Prior college GPA	3.38	3.43	0.668	2,385
Math placement score (std)	-0.080	0.057	0.438	5,478
ACT English	32.3	32.6	0.887	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	0.300	3,151
SAT Math	705	714	0.559	3,407
SAT Verbal	642	647	0.876	3,407
High school GPA	3.88	3.89	0.550	4,952
Took calculus in HS	0.814	0.838	0.428	5,104
Max parental education (omitted: less than high school)				
High school	0.071	0.070	0.273	5,641
Some college	0.064	0.051	0.411	
Bachelor's	0.253	0.241	0.433	5,641
Grad or professional degree	0.588	0.617	0.604	
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		0.836		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. All characteristics based on university administrative data.

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.[17]

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. 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 by regressing an indicator for post-intervention survey response on the full set of observed pre-treatment characteristics (Appendix Table A.6).[18] 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) had higher response rates, but the gender-by- performance interaction is not significant. Students with higher prior achievement, students in the statistics and engineering courses[19], engineering majors, younger students, and Asian students were also more likely to respond to the survey.

Survey response is independent of estimated treatment effects on course-taking 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 = 0.943). 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, reported below. 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, I lose

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

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

[19] Recall that instructors in these courses offered extra credit for both surveys.

precision but the point estimates are similar.

4 Empirical Method

I estimate the main effect of the intervention with the following specification:

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

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 indicator variables for all but one of of the 32 gender-by- course-by-above-median strata. (I also report estimates without covariates in the appendix.) 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).

To estimate effects by gender, I add in an interaction for female students:

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

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 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:

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

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 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 short-term measures of field specialization (course-taking in the semester following the intervention) based on administrative transcript data. I estimate effects on students' beliefs about their

relative performance using outcomes measured in the post-intervention survey. 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, presented in Appendix Table B.1 and described in more detail in Appendix B. Although they represent different conceptual approaches, the model- and randomization-based p-values produce virtually identical conclusions. To address concerns of data mining and the possibility of finding falsely significant results, I implement three types of adjustments for multiple hypothesis testing, reported in Appendix Table C.1. 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.

4.1 Outcome measures

My primary behavioral outcome is STEM persistence, operationalized as the number of credits attempted in the semester following the intervention, as well as an indicator for taking any STEM courses. I classify courses by two-digit Classification of Educational Program (CIP) code, developed and maintained by the U.S. Department of Education's National Center for Education Statistics.[20] These outcomes come from the administrative data; attrition or missingness occurs only if a student graduates or drops out.[21]

I measure 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 to their true percentile.[22] 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 average errors in beliefs; I report both the absolute value as well as a signed error to convey the direction of the error.

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

[20]The following subjects (CIP codes) are considered STEM: natural resources and conservation (03), computer and information sciences (11), engineering (14), biological and biomedical sciences (26), mathematics and statistics (27), physical sciences (40), and economics (45.06). I code economics (45.06) separately from the rest of the social sciences (45).

[21]If a student does not show up in the data in a given term, I code them as taking zero credits and courses. Fewer than two percent of control students do not appear in the data in the semester following the intervention.

[22]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.

in a STEM field; I can then compare their answers to the true median.[23] 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.

5 Results

5.1 Control Students' Beliefs about Relative Performance

To motivate the experimental results, I begin by describing students' beliefs in the absence of any intervention. In this section, I focus on control students only. I examine control students' beliefs at two points in time: at the beginning of the semester (generally in September) and again at the end of the semester (December). In my descriptive analyses of student beliefs, I limit the sample to control students who responded to both surveys to avoid any confounding changes due to differential response over time.

Control students begin the semester inaccurately predicting their performance.[24] The average control student overpredicts by 15.9 percentile ranks, meaning they expect to perform considerably better than they actually do. Because some students underpredict (a negative error), the average absolute value error is even larger in magnitude: 28 percentile ranks. There are significant differences by gender and performance. The average man assigned to the control condition overpredicts his final performance by 18.2 percentiles, while the average woman overpredicts by 13.7 ($p<0.05$). Low-performing (below-median) students tend to overestimate their performance (by 30.6 percentiles), while high-performing ones tend to underestimate, though to a lesser extent (average underprediction of 2.7 points).[25] Low-performing men are the most overconfident (overpredicting by an average of 34.4 percentiles, compared to 27.2 for low-performing women) while high-performing women are the most underconfident (underpredicting by 5.9 percentiles compared to less than a percentile for high-performing men). Panel (a) of Figure 2 visully summarizes the accuracy of these beginning-of-semester predictions by gender and realized performance.

Even absent intervention, we would expect students to update their beliefs over the course of the

[23]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 2015 through 2017 academic year and who declared a STEM major within three terms of taking the course.

[24]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 feedback.

[25]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.

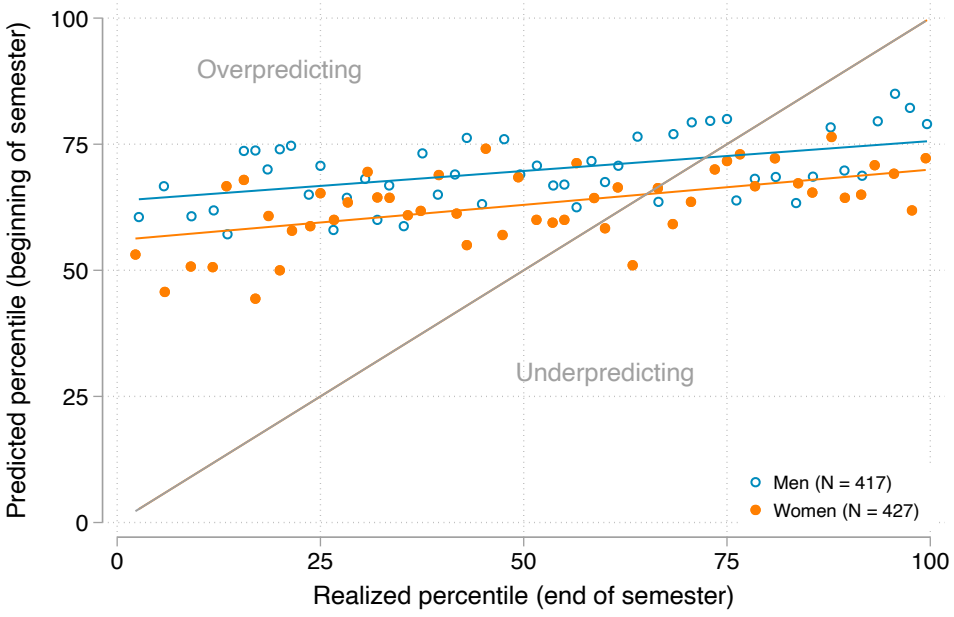
semester as they learn about their performance through exams, assignments, and other feedback. At the end of the semester (right before final exams), control students' predictions are more accurate than they were at the beginning. The average student still overpredicts, but by less: 5 percentiles compared to 15.9 at the start of the semester. Compared to an absolute value error of 28 percentiles at the beginning of the semester, the average control student's absolute error at the end of the semester is 19.2. The fact that the change in the signed error is similar to the change in the absolute value of the error suggests that it is primarily the students who were initially overpredicting who updated. Though both men and women have updated, a gender gap in beliefs remains: the average man assigned to the control condition overpredicts his final performance by 6.6 percentiles, while the average woman overpredicts by 3.4. The gender gap among low-performing students is only slightly smaller compared to the beginning of the semester: below-median men are 5.4 percentiles more overconfident than women (15.2 vs. 9.8). The gender gap among high-performing students has shrunk to 2.7 percentile points and is not statistically significant. These changes are reflected in Panel (b) of Figure 2.

I next turn to what students believe about the performance of 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. 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.

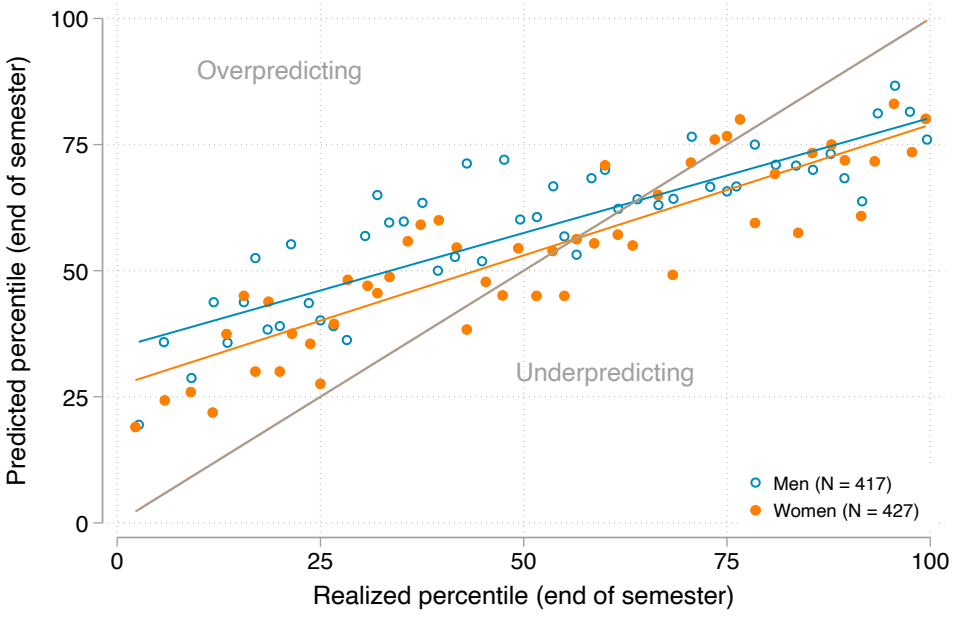
Control students' beliefs about this median change little over the semester (Figure 3, Panel (b)). This is unsurprising; though they learn about their own performance and, to a lesser extent, that of their peers, they receive no direct information about STEM majors' grades in particular. By December, when they respond to the post-intervention survey, 26 percent of control men and 17 percent of 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).[26]

[26]Students also responded to questions about their beliefs on the overall course median (for all students) 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

Figure 2: Control Student Beliefs about Own Percentile by Gender



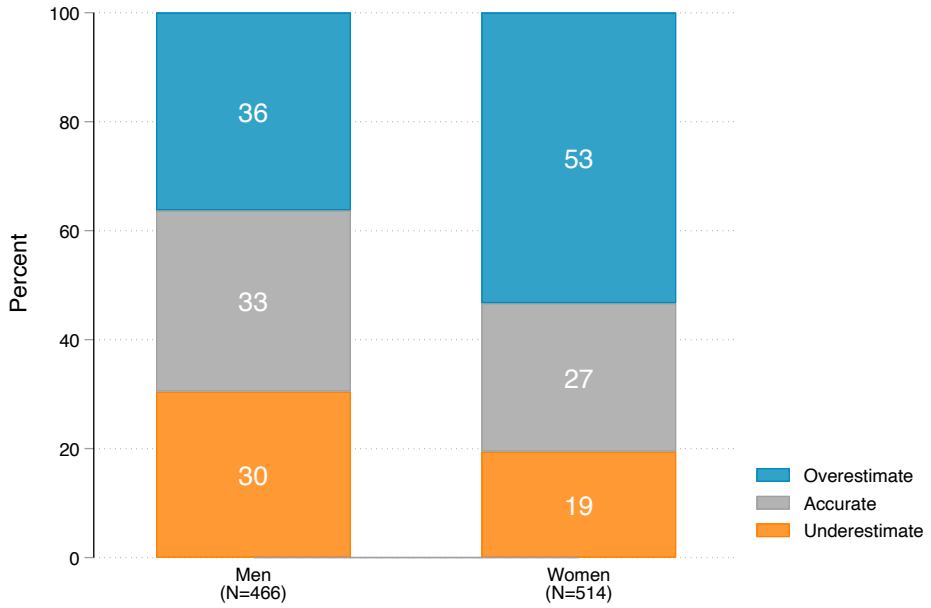
(a) Beginning of Semester Beliefs



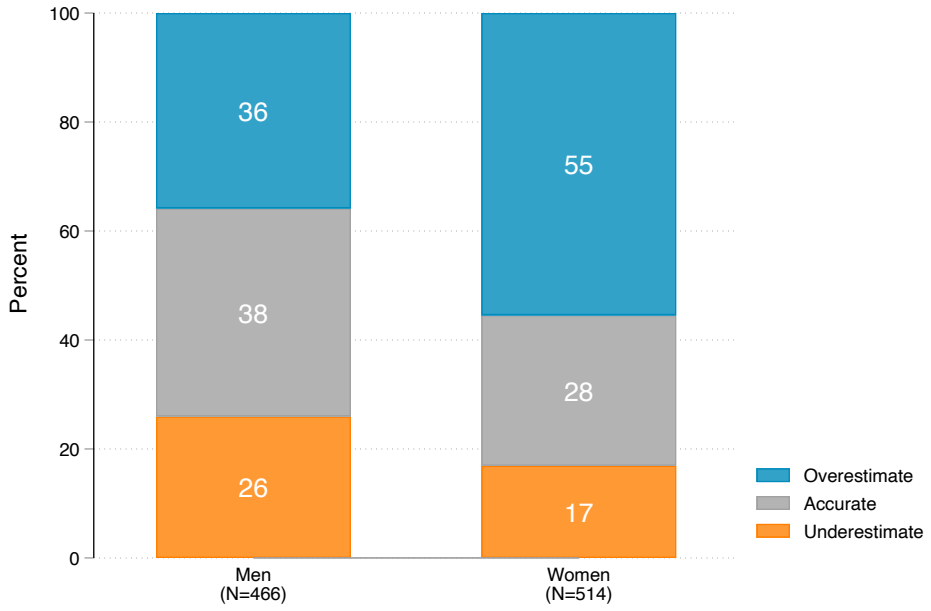
(b) End of Semester Beliefs

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

Figure 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.

The 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. (A single STEM course is usually four credits, so this represents half of a course.) Though consistent with gender differences in confidence explaining gaps in persistence, this relationship is correlational and does not account for the myriad factors which may differ by gender. To isolate the causal role of relative performance beliefs, my experiment aims to exogenously change beliefs and study how academic decisions change as a result.

5.2 Effect of Intervention on STEM Persistence

Table 2 reports my primary experimental results: estimated treatment effects on STEM persistence, measured as STEM credits taken in the semester following the intervention.[27] 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 ($p<0.1$), a decrease of two percent relative to the control mean of 8.5. The second two columns estimate effects by gender and reveal that 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 0.08 (one percent). The -0.28 credit effect for men would be equivalent to roughly 120 men, or seven percent of treated male students, taking one fewer four-credit STEM course. 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. These changes are consistent with men correcting their overconfidence and taking fewer STEM courses as a result; I investigate this at length in the remainder of the paper.

I find a small 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 point estimates for men and women are identical to three digits and statistically indistinguishable.

For high-performing students, who were eligible for the second treatment arm, I test for differential effects on STEM course-taking by treatment arm (Appendix Table A.9) but find none, for women or men.[28]

about the selection into STEM, with women setting the bar for STEM higher.

[27] Treatment effects on STEM course-taking outcomes estimated without covariates are included as Appendix Table A.8. The results are very similar.

[28] I designed a three-armed experiment assuming I would have two semesters of students in my sample. The cancellation of the second round due to the pandemic left me with half of my planned sample size and less statistical

Since I find no evidence of a differential treatment effect, for the remainder of the paper I combine the treatment arms and consider the effect of receiving any type of informational treatment. Recall that 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.

Table 2: 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 (Equation 1). Treatment effects by gender estimated from a single regression of the outcome on assignment to either treatment, female, and treatment-times-female, controlling for student academic and demographic characteristics and randomization strata dummies (Equation 2). Robust standard errors reported. Course-taking outcomes based on University of Michigan administrative data.

5.3 Effect of Intervention on Student Beliefs

The intervention aimed to change students' behavior by correcting their beliefs about their relative performance. I estimate treatment effects on students' beliefs using survey measures of relative performance beliefs similar to those described in Section 5.1. The first measures the accuracy of students' beliefs about their own relative performance by subtracting the student's true percentile from what they estimate their percentile to be at the end of the semester. Here, I use mid-semester performance as the realized percentile, because end-of-semester performance could itself be affected by the intervention if students adjust their effort. (For this reason, the control means in the treatment effects tables differ from the values reported in Section 5.1.) I test for effects on performance directly in Section 5.5.[29] I report both an absolute value

power to distinguish between treatment arms.

[29] I also estimate effects on a version of the percentile belief outcomes using final performance rather than mid-semester performance as the realized performance (Appendix Table A.10). The signs are similar but the magnitudes somewhat smaller. This is not surprising given that the intervention told students their mid-semester

measure as well as a signed measure that captures the direction of the error. Second, I measure the accuracy of beliefs about the performance of STEM majors by creating two indicator variables for whether a student is over- or underestimating the course median for students who go on to major in STEM.

Table 3 shows treatment effects on beliefs outcomes, for the full sample as well as separately for men and women.[30] Effects on the absolute value of the error in 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. This can be seen in Panel (a) of Figure 4, which shows that both over- and underconfident men update their beliefs as a result of the treatment. This is reflected by the line through the treated points shifting closer to the 45-degree line, relative to the control men. For women, on the other hand, the treated and control trends are indistinguishable, showing that the treatment did not cause women to update their beliefs about their percentile rank, on average.

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 (bottom panel of Table 3). Receiving the informational intervention made men 5.2 percentage points less likely to underestimate the median and made women 5.1 percentage points less likely to overestimate. The gender gap in underestimating among control students is 9.8 percentage points (with men more likely to underestimate) and the control gap in overestimating is 17.7 percentage points (with women more likely to overestimate). Comparing control and treatment gender gaps, the treatment closes the gap in both measures by roughly a third. Both changes suggest that men are becoming less overconfident relative to women and are broadly consistent with the observed changes to men's behavior, with men taking fewer STEM credits as a result of receiving information. Although women did correct their beliefs (at least about the STEM median), they did not change their behavior in response.[31]

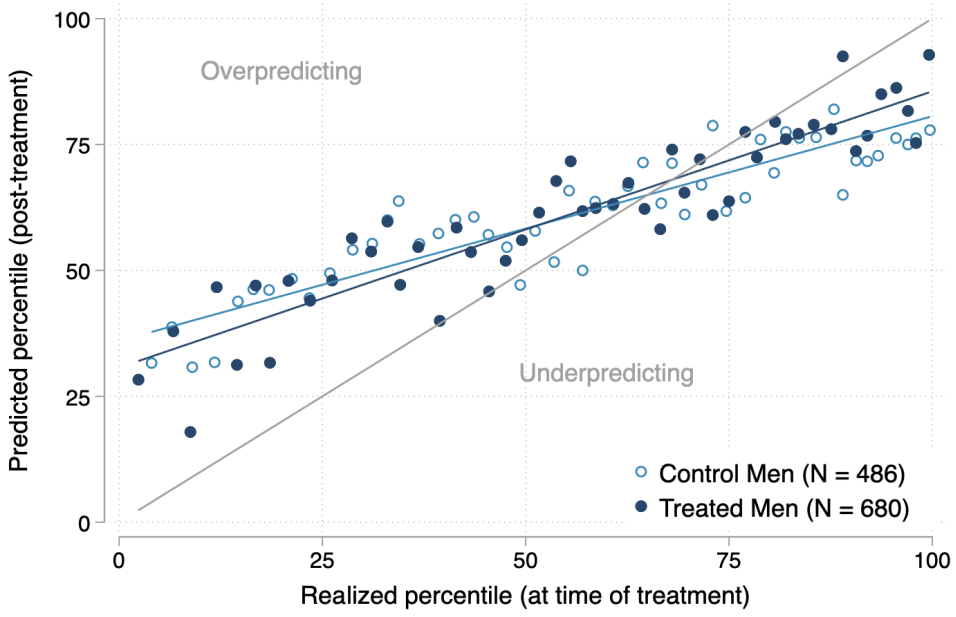
---

percentile; they updated their beliefs in the direction of the signal they received.

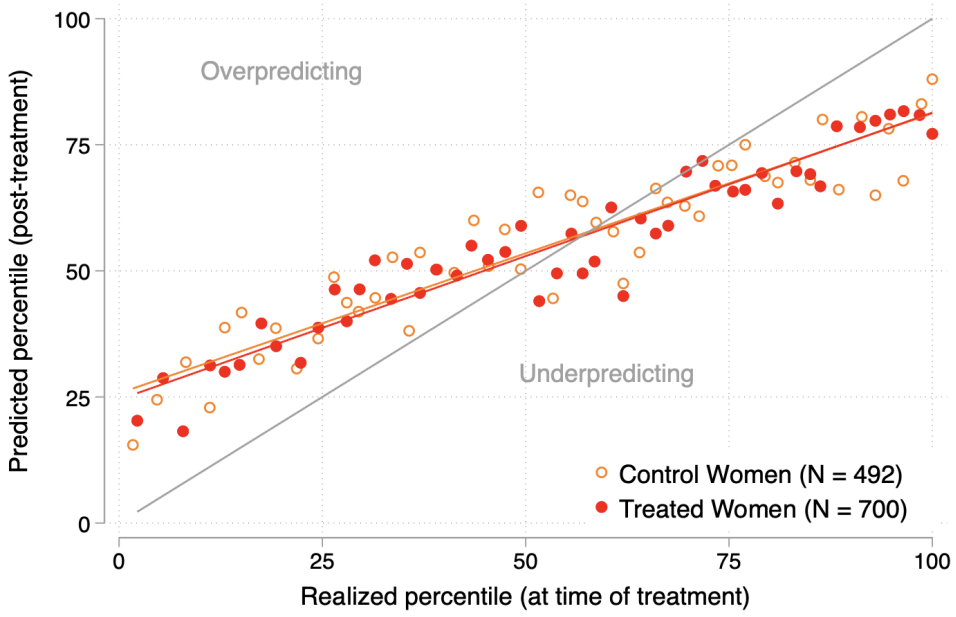
[30] Treatment effects on beliefs outcomes estimated without covariates are included as Appendix Table A.11. The results are very similar.

[31] As a robustness check, I estimate treatment effects on STEM course-taking outcomes but limit my sample to

Figure 4: Post-Treatment Student Beliefs about Own Percentile, by Treatment Status and Gender



(a) Men



(b) Women

Notes: The *x*-axis measures students' realized percentile within the course, measured at the time of the intervention. This corresponds to the percentile students were informed of as part of the intervention. The *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.

Table 3: Estimated Effect of Intervention on Students' Beliefs about Themselves and Other STEM Majors, Overall and by Gender

	Absolute value of error in percentile beliefs (| Predicted – realized |)			Signed error in percentile beliefs (Predicted – realized)		
	All	Men	Women	All	Men	Women
Treatment effect	-1.485**	-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 (Equation 1). Treatment effects by gender estimated from a single regression of the outcome on assignment to either treatment, female, and treatment-times-female, controlling for student academic and demographic characteristics and randomization strata dummies (Equation 2). Robust standard errors reported. All beliefs outcomes based on response to post-intervention survey. Realized performance is measured mid-semester, at the time of intervention.

5.4 Heterogeneity by Pre-Intervention Beliefs and Performance

To better understand how and for which students the intervention affected behavior, I estimate heterogeneous treatment effects based on students' initial beliefs in the pre-intervention survey. We would expect that the

students who responded to the post-intervention survey. The results, shown in Appendix Table A.12, produce very similar point estimates to Table 2, although they are less precise. 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.13. The point estimates are similar to the ones in Table 3, though again less precise. 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.

men who respond by taking fewer STEM courses are those who were initially overconfident about their relative performance and for whom the informational intervention contained bad news. Since I have two measures of beliefs, I examine two types of prior belief heterogeneity.

First, I categorize students' initial beliefs about their percentile by whether they were initially underpredicting their percentile (meaning they received good news), or initially overpredicting (meaning they received bad news).[32] It is important to note that initial beliefs are measured in September, and the treatment tells students their percentile as of November. I do not observe their beliefs at the precise time of treatment. It is likely that students have updated in the first half of the semester, which could mute estimated heterogeneity by initial beliefs. Furthermore, results relying on survey data have considerable missingness, so should be interpreted with caution.

The results, shown in Panel A of Table 4, are not wholly consistent with a belief updating story. The effect for students who were initially overpredicting their percentile, and who therefore received bad news, is negative, as expected: -0.241 credits. However, the effect is only marginally statistically significant ($p<0.1$). It is also not significantly different from the effect for students who received good news about their percentile, which is also negative, but insignificant (-0.144 credits). Disaggregating by gender, it was the men who were initially *under*confident and received *good* news about their performance who decreased their STEM course-taking, by 0.55 credits or 5 percent. I do not detect any change for initially overconfident men, or for any subgroup of women.

In Panel B of Table 4, I group students by the accuracy of their initial beliefs about the median course grade for STEM majors: initially accurate, initially overestimating the median, or initially underestimating. A student who was initially overestimating the median would have received good news, since their own relative position is better than they thought. These results also do not show expected patterns. A simple belief updating model would predict no change in behavior for those who were initially correct, a positive effect for those who received good news, and a negative effect for those who received bad news about the median. In fact, the effect for students who received good news (those who initially overestimated the median) is negative and marginally significant (-0.296 credits, $p<0.1$). Although the effect for initially correct students is not statistically significant, it is nearly identical to the effect for those who received good news (-0.297). Students who received bad news about the median did not change their course-taking (effect of 0.036 credits).

[32] I compare students' prediction of their percentile, which they make at the beginning of the semester, to their percentile at the time of the intervention, mid-semester. The mid-semester percentile is what treated students are told as part of the intervention.The small number of students who accurately predict their percentile (N=43) are grouped with those who underpredict.

By gender, men who correctly identified the median at the beginning of the semester appeared to change their behavior the most, reducing STEM course-taking by 0.6 credits (6 percent). Men who initially overestimated the median—meaning they were relatively underconfident about their position relative to others, and received good news—also decreased their course-taking by 0.43 credits ($p<0.1$). The men we would expect to react negatively, those who were initially underestimating the median (and therefore overestimating their own relative position) did not change their behavior. Again, no subgroup of women changed their course-taking.

One possible explanation for this somewhat puzzling pattern in behavior is that subgroups of students did not update their beliefs in the correct direction. I explore heterogeneity in belief updating to try to reconcile these patterns. In Appendix Table A.14, I estimate heterogeneous treatment effects on beliefs, by initial beliefs. While some of the results imply that students correctly updated (e.g., men who were initially underestimating the median adjusted that belief downwards), others do not, and the patterns do not generally match up with the effects on course-taking. However, these results rely on students who responded to both surveys, and are not a representative sample.

I also examine heterogeneity by student performance. We might expect information highlighting academic performance to discourage low-performing students and encourage higher-performing students to take more STEM. Table 5 does confirm that the lowest-performing students (those in the bottom quartile of their class at the time of the intervention) react most negatively, decreasing their STEM course-taking by 0.475 credits (6.5 percent, $p<0.05$). The effects by gender reveal that this is driven by low-performing men, who see a decrease of 0.665 STEM credits (8 percent). However, the highest-performing men (those in the top quartile) decrease their STEM course-taking by a similar magnitude: 0.545 credits or 5 percent. This last result is again at odds with a belief updating story, since the highest performing men are somewhat underconfident, on average (see Figure 2). I detect no significant change for any subgroup of women.

In Appendix Table A.15, I examine belief updating by pre-treatment performance. Lower-performing men (those in the bottom two quartiles) update their beliefs in a way that means they are becoming less relatively overconfident. They correct their beliefs about their own percentile and the STEM median. The lowest-performing (bottom quartile) men seem to *over*correct their beliefs about the median grade for STEM majors; the decrease in underestimating the median is matched by a similarly-sized increase in overestimating. All of these changes are consistent with low-performing men correcting (or overcorrecting) their relative overconfidence and being discouraged from STEM. The highest performing men correctly update their beliefs about their percentile upwards, by nearly 5 points on average. This suggests they are becoming less underconfident, which does not seem to explain why this group also takes *fewer* STEM

Table 4: Estimated Effect of Intervention on Students' STEM Course-taking, by Pre-Intervention Beliefs

	Number of STEM credits one semester post intervention		
	All	Men	Women
A. Treatment effect by own percentile beliefs			
Students underpredicting	-0.144	-0.550**	0.213
percentile (got good news)	(0.196)	(0.268)	(0.282)
	[9.060]	[10.373]	[7.916]
Students overpredicting	-0.241*	-0.177	-0.310
percentile (got bad news)	(0.146)	(0.202)	(0.211)
	[8.278]	[9.187]	[7.279]
N	3,664	1,874	1,790
B. Treatment effect by STEM median beliefs			
Students who correctly	-0.297	-0.590**	0.039
identified STEM median	(0.218)	(0.293)	(0.325)
	[8.562]	[9.506]	[7.487]
Students initially overestimating	-0.296*	-0.430*	-0.199
median (got good news)	(0.164)	(0.254)	(0.216)
	[8.011]	[9.207]	[7.152]
Students initially underestimating	0.036	0.080	-0.042
median (got bad news)	(0.220)	(0.269)	(0.379)
	[9.384]	[9.925]	[8.482]
N	3,915	1,973	1,942

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, indicators for pre-intervention beliefs, and treatment-by-pre-beliefs interactions, controlling for student academic and demographic characteristics and randomization strata dummies. Treatment effects by gender estimated from a single regression with a three-way interaction between treatment, female, and pre-intervention beliefs, controlling for student academic and demographic characteristics and randomization strata dummies. Pre-intervention beliefs are based on responses to the pre-intervention survey. In Panel A, underpredicting means the student's self-prediction of their percentile was lower than (or equal to) the percentile the intervention informed them of, while overpredicting means their self-prediction was higher than the information they received. In Panel B, students are categorized by whether they initially correctly identified the course median for students who go on to major in STEM. Robust standard errors reported. Control means in square brackets. Course-taking outcomes based on University of Michigan administrative data.

credits. High-performing women correct their overestimation of the STEM median by 10 percentage points; this implies an improvement to their relative confidence, since the bar is not as high as they thought.

Taken together, the estimated effects of the informational intervention on students' beliefs and subsequent behavior are suggestive that men's overly confident beliefs about their relative performance are partially responsible for their higher rates of STEM persistence. By inducing them to accurately revise their beliefs about their relative performance, the experiment caused men to take fewer STEM credits. Women, on the other hand, revised their beliefs in a direction that should make them less underconfident about their relative performance, but did not change their behavior. Patterns by pre-intervention performance suggest that initially low-performing men—who tend to be overconfident—were discouraged by information about relative performance, implying that male overconfidence rather than female underconfidence appears to be a determinant of the gender gap in field specialization. The behavior of higher-performing, underconfident men—who also decrease their STEM course-taking—presents a puzzle, which I discuss further in Section 6.

5.5 Supplemental Outcomes and Heterogeneity

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 important for educators who care about improving performance, and could also be a 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 effort if they realize their performance is not adequate for a STEM major.

I estimate treatment effects on two performance outcomes: final exam and final course scores, both measured as percent scores out of 100 (Table 6).[33] 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

[33]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 exam lowest score, so many choose not to take the final), so I do not include it in my analyses of final exam performance.

Table 5: Estimated Effect of Intervention on Students' STEM Course-taking, by Pre-Intervention Performance

	Number of STEM credits one semester post intervention		
	All	Men	Women
Treatment effect by pre-treatment performance			
Students in bottom quartile	-0.475**	-0.665**	-0.276
at time of treatment	(0.198)	(0.280)	(0.282)
	[7.342]	[8.408]	[6.240]
Students in second quartile	-0.004	0.015	-0.024
	(0.190)	(0.264)	(0.272)
	[8.320]	[9.115]	[7.492]
Students in third quartile	-0.044	0.130	-0.234
	(0.185)	(0.251)	(0.272)
	[9.048]	[9.803]	[8.214]
Students in top quartile	-0.196	-0.545**	0.233
	(0.183)	(0.229)	(0.294)
	[9.915]	[11.086]	[8.493]
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, indicators for pre-intervention performance quartile, and treatment-by-performance-quartile interactions, controlling for student academic and demographic characteristics and randomization strata dummies. Treatment effects by gender estimated from a single regression with a three-way interaction between treatment, female, and pre-intervention performance quartile, controlling for student academic and demographic characteristics and randomization strata dummies. Pre-intervention performance is measured at the time of treatment, in November. Robust standard errors reported. Control means in square brackets. Course-taking outcomes based on University of Michigan administrative data.

performance by less than a percentage point, suggesting effort and performance were not a key mechanism through which changing beliefs affected behavior.

The intervention could change students' beliefs about their ability to succeed in STEM, which could serve as an intermediate channel between their beliefs about their performance and their behavior. To measure this, I construct an index capturing students' beliefs about their ability to succeed in STEM, which aggregates responses to items about their grades being "good enough" for STEM, a series of STEM-self-efficacy items, and items about identifying with being a "math person" or "science person".[34] The results are included as the bottom left 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.

Finally, by calling attention to grades and academic performance, the intervention may have increased students' academic stress levels, a possible mechanism to explain the negative effects on STEM course-taking for men or the lack of effect for women. To test this, I estimate treatment effects on a subjective measure of grade stress: a standardized version of an item asking students to rate their general stress and anxiety level about their academic performance and grades. The bottom right panel of Table 6 shows no change to students' stress about grades, overall or by gender.

Appendix Tables A.16 through A.18 report estimated effects on STEM persistence by a number of pre-treatment characteristics, including student level, intended major, course subject, and gender composition of the course. The heterogeneity results imply that students who we would expect to be on the margin of specializing in STEM—younger students and students already interested in STEM—are the ones who change their course-taking behavior (Appendix Table A.16). However, I lack the statistical power to reject equality in effects across groups.

In terms of course subject, I find that students in the computer science and statistics courses decreased their STEM course-taking by the most (Appendix Table A.17). However, by splitting the sample into seven subjects, I don't have the power for subject-by-subject comparisons. The p-value on a test for whether course subject jointly predicts the treatment effect is 0.08. I find no significant differences by gender composition of the course (Appendix Table A.18), though this analysis is again underpowered due to the loss of sample size from the cancellation of the second round of the study.

[34]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.

Table 6: Estimated Effect of Intervention on Students' Performance, Beliefs about Ability to Succeed in STEM, and Stress About Grades

	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)			Grade stress (std. dev. units)		
	All	Men	Women	All	Men	Women
Treatment effect	0.024	0.013	0.035	0.001	-0.029	0.029
	(0.025)	(0.035)	(0.035)	(0.039)	(0.058)	(0.051)
P-value, women vs. men			0.656			0.451
Control mean	0	0.116	-0.108	0	-0.239	0.221
N	2,687	1,317	1,370	2,638	1,290	1,348

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 (Equation 1). Treatment effects by gender estimated from a single regression of the outcome on assignment to either treatment, female, and treatment-times-female, controlling for student academic and demographic characteristics and randomization strata dummies (Equation 2). 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. Grade stress is based on a post-intervention survey item asking students to rank the stress and anxiety they feel about academic performance and grades.

5.6 Major Choice and Predicted Long-term Outcomes

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 Table 7, 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.[35]

I ultimately am interested in students' choice of college major. However, because the studied students are still early in their academic careers, this outcome does not yet exist. While course-taking is a short-term proxy for and important precursor to major choice, I also use information from the survey and the effects on course-taking to speculate on the choice of a STEM major.

I examine subjective interest in STEM in two ways. The first is simply whether a student 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 general interest in STEM, intention to seek academic advising in a STEM field, and intention to take subsequent STEM courses.[36] As shown in Table 8, I find small, negative, statistically insignificant effects on subjective STEM intent and small negative effects on STEM interest. Although the effects on the STEM interest index are negative for both men and women (-0.045 and -0.085 standard deviations, respectively), if anything, the effect is more negative for women, which does not align with STEM course-taking effects. However, both effects are small (less than one tenth of a standard deviation) and I cannot reject that they're equal.

I also estimate treatment effects on students' predicted STEM degree receipt, following 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,

[35] 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.

[36] Like with the STEM success index, the construction of the interest index follows Kling et al. (2007).

Table 7: Estimated Effect of Intervention on Number of Credits in Non-STEM Subjects

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

Notes: $*p < 0.1$; $**p < 0.05$; $***p < 0.01$. Treatment effects for all students estimated from a regression of the outcome on assignment to either treatment, controlling for student academic and demographic characteristics and randomization strata dummies (Equation 1). Treatment effects by gender estimated from a single regression of the outcome on assignment to either treatment, female, and treatment-times-female, controlling for student academic and demographic characteristics and randomization strata dummies (Equation 2). Robust standard errors reported. Course-taking outcomes based on University of Michigan administrative data and measured in the semester following the intervention. "Social science" excludes economics and 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 8: Estimated Effects of Intervention on Students' Subjective Interest in STEM and Predicted Degree Receipt

	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(Equation 1). Treatment effects by gender estimated from a single regression of the outcome on assignment to either treatment, female, and treatment-times-female, controlling for student academic and demographic characteristics and randomization strata dummies (Equation 2). 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.

with some assumptions, provides an unbiased estimate of the ATE on the long-term outcome.[37] The bottom panel of Table 8 shows estimates for treatment effects on predicted long-term degree. The estimated effects

[37]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.

for all students as well as for men and women are small, negative, and not statistically significantly different from zero.

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

6 Discussion

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

An implication of these models is that, assuming a positive relationship between beliefs about major-specific ability and the expected payoff to a major, those performing better in STEM than they expected should be (weakly) more likely to pursue STEM, while those who receive a negative signal should be 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 intervention—is how those beliefs are formed and why they persist. One possibility is that students are incorporating signals from other sources like standardized test scores and previous coursework, and men have received signals that are more positive than women. I can investigate this in the data, and while men are more likely to have taken calculus in high school and have higher quantitative test scores, controlling for all of these factors does not change the gender gap in beliefs. Theory paired with lab-based studies of belief updating suggest that exaggerated stereotypes about groups (e.g., men are much better at quantitative subjects) can persist despite very small true differences, due to people using mental shortcuts to make predictions about themselves or others (Bordalo et al., 2016). This would explain men overestimating and women underestimating their own quantitative ability.

Consistent with field-specific beliefs mattering for specialization, men updating relative beliefs downwards leads to them taking fewer STEM credits. One puzzling effect of my informational intervention was to decrease STEM persistence not just for low-performing, overconfident men, but for high-performing, initially underconfident men, as well. One possible explanation is that these high-performing men came to view STEM as less difficult and therefore less prestigious, which decreased their interest. However, I lack the data to formally test this hypothesis.

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 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.

7 Conclusion

Gender differences in college field specialization and their implications for the labor market are of great interest to policymakers. There is a strong theoretical and empirical basis for believing that gender differences in 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. In a field experiment across seven introductory STEM courses, I provided 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 theory that beliefs matter for educational choices, 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. These findings advance our understanding of how beliefs factor into academic decisions. Prior work has disagreed on whether female underconfidence versus 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).

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. Men leaving STEM could also have 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 others. The changing gender and achievement composition of students in STEM courses 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. The degree of overconfidence among the students in my sample may be 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 treatment effects I find 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 set of students. A more intensive 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. However, increasing women's STEM participation likely requires additional approaches.

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.

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.

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.

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.

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.

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.

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.

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.

University of Michigan Center for Academic Innovation (2019). ECoach survey data.

University of Michigan Office of Enrollment Management (2021). Learning analytics data architecture (LARC) data set.

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.

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

Appendix

Appendix A. Supplemental Figures and Tables

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

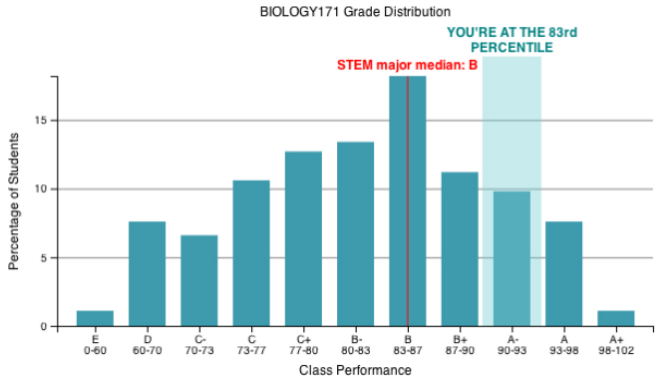
Your Bio 171 grade | And your major

[redacted],

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

HERE'S HOW YOU'RE DOING.

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



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

HERE'S HOW GRADES OFTEN LOOK.

The typical median grade for BIOLOGY 171 is:

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

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

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

AS YOU PLAN YOUR SCHEDULE...

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

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

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

You can view course options for Winter 2020 here.

~ The ECoach Team

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

Your Bio 171 grade | And your major

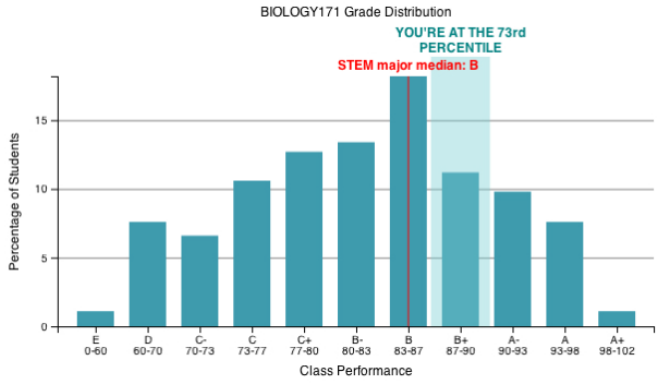
[redacted]

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

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

YOU'RE PERFORMING LIKE A STEM MAJOR!

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



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

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

HERE'S HOW GRADES OFTEN LOOK.

The typical median grade for BIOLOGY 171 is:

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

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

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

AS YOU PLAN YOUR SCHEDULE...

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

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

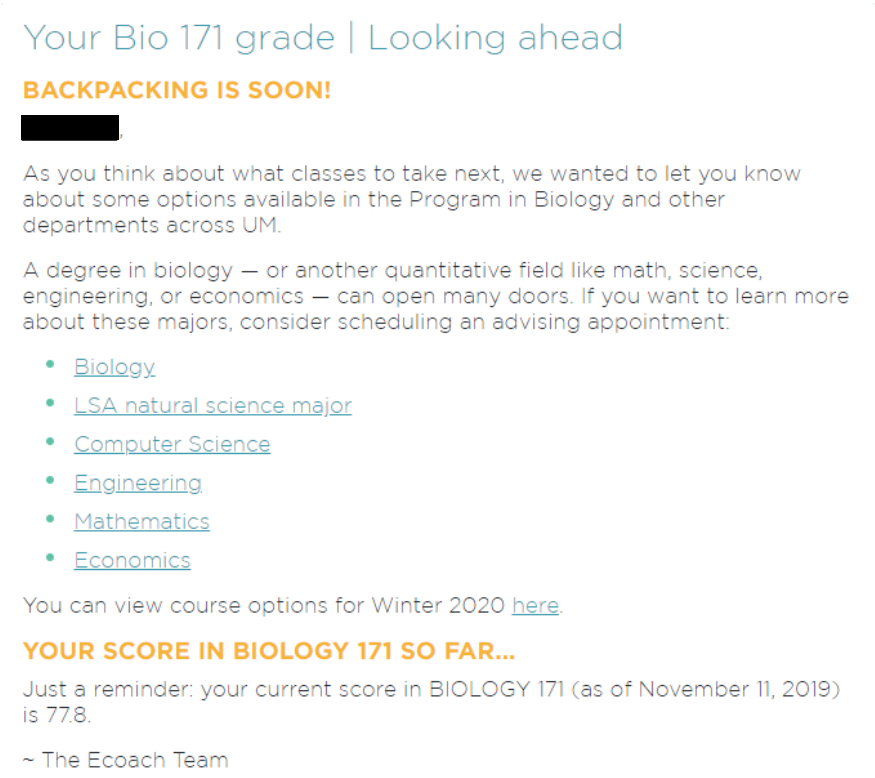
- Biology
- LSA natural science major
- Computer Science
- Engineering
- Mathematics
- Economics

You can view course options for Winter 2020 here.

Congrats again — keep up the good work!

~ The ECoach Team

Figure A.3: Sample Intervention Message: Control Group



Your Bio 171 grade | Looking ahead

BACKPACKING IS SOON!

[redacted],

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

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

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

You can view course options for Winter 2020 here.

YOUR SCORE IN BIOLOGY 171 SO FAR...

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

~ The Ecoach Team

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

	Control	Info-only	Info + encour.	p-value
Female	0.461	0.459	0.461	-
Class standing (omitted: senior)				
First year	0.418	0.420	0.404	0.725
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	0.180
Hispanic	0.041	0.055	0.044	0.305
Asian	0.319	0.343	0.330	0.493
Black	0.013	0.007	0.014	0.174
Declared major (omitted: other)				
Undeclared	0.545	0.541	0.539	0.964
Engineering	0.260	0.255	0.266	0.709
Math, science, or economics	0.104	0.112	0.091	0.306
Academic and demographic characteristics				
In-state	0.480	0.460	0.490	0.410
Prior college GPA	3.61	3.61	3.63	0.807
Math placement score (std)	0.330	0.365	0.331	0.540
ACT English	33.4	33.3	33.5	0.374
ACT Math	32.3	32.3	32.4	0.815
ACT Reading	32.7	32.3	32.7	0.057
ACT Science	32.2	32.1	32.2	0.896
SAT Math	738	739	735	0.306
SAT Verbal	661	659	661	0.902
High school GPA	3.92	3.92	3.91	0.629
Took calculus in HS	0.873	0.882	0.858	0.313
Max parental education (omitted: less than high school)				
High school	0.042	0.055	0.040	0.284
Some college	0.038	0.029	0.037	0.499
Bachelor's	0.242	0.221	0.248	0.366
Grad or professional degree	0.669	0.683	0.663	0.620
Family income (omitted: less than $50,000)				
$50,000-100,000	0.158	0.170	0.166	0.803
Above $100,000	0.731	0.704	0.716	0.502
Total N	940	943	940	2,823

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

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

	Men			Women		
	Control	Treat	p-value	Control	Treat	p-value
Class standing (omitted: senior)						
First year	0.446	0.407	0.078	0.419	0.428	0.688
Sophomore	0.370	0.405	0.236	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	0.475	0.556	0.544	0.380
Hispanic	0.078	0.072	0.875	0.062	0.064	0.303
Asian	0.258	0.300	0.201	0.248	0.277	0.482
Black	0.025	0.018	0.672	0.052	0.033	0.212
Declared major (omitted: other)						
Undeclared	0.487	0.477	0.947	0.638	0.650	0.415
Engineering	0.305	0.314	0.842	0.153	0.149	0.384
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	0.366
Prior college GPA	3.30	3.37	0.812	3.44	3.48	0.365
Math placement score (std)	0.080	0.242	0.081	-0.251	-0.146	0.564
ACT English	32.4	32.5	0.287	32.2	32.7	0.397
ACT Math	31.9	32.1	0.641	29.8	30.4	0.663
ACT Reading	32.0	31.8	0.026	32.0	31.9	0.105
ACT Science	31.6	31.8	0.464	30.1	30.4	0.464
SAT Math	717	730	0.133	690	694	0.019
SAT Verbal	646	654	0.298	638	639	0.159
High school GPA	3.87	3.88	0.688	3.90	3.90	0.651
Took calculus in HS	0.832	0.867	0.097	0.796	0.806	0.653
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	0.591	0.077	0.061	0.534
Bachelor's	0.242	0.237	0.973	0.265	0.245	0.276
Grad or professional degree	0.612	0.639	0.647	0.561	0.593	0.785
Family income (omitted: less than $50,000)						
$50,000-100,000	0.175	0.185	0.307	0.190	0.195	0.462
Above $100,000	0.658	0.664	0.390	0.588	0.619	0.990
P-value on F-test of all X's		0.817			0.623	
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 Only

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

Continued on next page

Table A.4 – *Continued from previous page*

Characteristic	Viewed message
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: 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
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
Grade stress	0.462	2,638
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	Characteristic	Took survey
Female	0.071***	*Declared major (omitted: other)*	
	(0.017)	Undeclared	0.006
Above course median	0.070***		(0.019)
	(0.017)	Engineering	0.080***
Female*above median	-0.022		(0.025)
	(0.022)	Math, science, or econ	0.031
Course (omitted: Econ section 1)			(0.027)
Biology	0.561***	*Acad. and demog. characteristics*	
	(0.024)	In state	0.009
Chemistry	0.017		(0.012)
	(0.017)	Prior college GPA	0.109***
Computer Science	0.485***		(0.020)
	(0.022)	College GPA missing	0.418***
Engineering	0.642***		(0.071)
	(0.027)	Math placement score	0.002
Physics	0.086***		(0.002)
	(0.027)	Placement score missing	-0.007
Statistics	0.641***		(0.048)
	(0.017)	ACT English	0.001
Econ (section 2)	0.610***		(0.003)
	(0.028)	ACT math	-0.001
Class standing (omitted: senior)			(0.003)
First year	0.080**	ACT reading	0.000
	(0.035)		(0.003)
Sophomore	0.086***	ACT science	-0.005*
	(0.031)		(0.003)
Junior	0.023	ACT missing	-0.168*
	(0.031)		(0.093)
Race/ethnicity (omitted: other/multiple)		SAT math	-0.000
White	0.007		(0.000)
	(0.022)	SAT verbal	-0.000***
Hispanic	0.008		(0.000)
	(0.030)	SAT missing	-0.295***
Asian	0.067***		(0.104)
	(0.024)	HS GPA	0.123**
Black	-0.032		(0.053)
	(0.039)	HS GPA missing	0.479**
Race/ethnicity missing	0.052		(0.207)
	(0.039)	Took calculus in HS	-0.001
			(0.017)
		HS calculus missing	-0.016
			(0.026)

Continued on next page

Table A.6 – *Continued from previous page*

Characteristic	Took survey
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: Table shows coefficients and robust 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	0.310	2,784
Sophomore	0.417	0.428	0.900	
Junior	0.129	0.136	0.340	
Race/ethnicity (omitted: American Indian or multiple race/ethnicities)				
White	0.533	0.535	0.916	2,698
Hispanic	0.061	0.063	0.194	
Asian	0.304	0.317	0.640	
Black	0.030	0.019	0.268	
Declared major (omitted: other)				
Undeclared	0.601	0.574	0.254	2,784
Engineering	0.201	0.209	0.300	
Math, science, or economics	0.095	0.104	0.502	
Academic and demographic characteristics				
In-state	0.506	0.517	0.291	2,784
Prior college GPA	3.44	3.47	0.204	1,172
Math placement score (std)	-0.025	0.107	0.869	2,676
ACT English	32.5	32.7	0.502	1,567
ACT Math	30.9	31.4	0.814	1,567
ACT Reading	32.1	31.9	0.008	1,567
ACT Science	30.9	31.1	0.367	1,567
SAT Math	708	717	0.251	1,623
SAT Verbal	640	647	0.815	1,623
High school GPA	3.89	3.90	0.999	2,374
Took calculus in HS	0.817	0.842	0.719	2,506
Max parental education (omitted: less than high school)				
High school	0.069	0.066	0.386	2,746
Some college	0.061	0.049	0.581	
Bachelor's	0.255	0.241	0.377	2,746
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		0.943		2,784
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: 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	0.910	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 only for randomization strata dummies. Treatment effects by gender estimated from a single regression of the outcome on assignment to either treatment, female, and treatment-times-female, controlling only for randomization strata dummies. Estimates with covariates are reported in Table 2. Robust standard errors reported. Course-taking outcomes based on University of Michigan administrative data.

Table A.9: Estimated Effect of Intervention on Students' STEM Course-taking by Gender and Treatment Arm, Above-Median Students Only

	Number of STEM credits one semester post intervention			Took any STEM courses one semester post intervention		
	All	Men	Women	All	Men	Women
Pooled effect	-0.151	-0.285*	0.007	-0.011	-0.012	-0.009
	(0.131)	(0.171)	(0.202)	(0.008)	(0.009)	(0.014)
P-value, women vs. men			0.271			0.855
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: $*p < 0.1$; $**p < 0.05$; $***p < 0.01$. All effects in this table are estimated on the sample of above-median students only. 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 the outcome on an indicator for receiving either treatment (Equation 1). To estimate pooled effects separately for above-median men and women, an interaction between any treatment and female is added (Equation 2). Treatment effects of the information-only and info-plus-encouragement intervention for above-median students are estimated using the same specifications as above, but with two separate treatment indicators (Equation 3). 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.10: Estimated Effect of Intervention on Students' Beliefs about Themselves, Comparing Beliefs to End of Semester Performance

	Absolute value of error in percentile beliefs (| Predicted - realized |)			Signed error in percentile beliefs (Predicted - realized)		
	All	Men	Women	All	Men	Women
Treatment effect	-1.381**	-1.466	-1.298	0.329	0.086	0.567
	(0.651)	(0.987)	(0.861)	(0.929)	(1.382)	(1.250)
P-value, women vs. men			0.898			0.797
Control mean	19.351	20.241	18.469	4.952	7.2	2.722
N	2,355	1,166	1,189	2,355	1,166	1,189

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 (Equation 1). Treatment effects by gender estimated from a single regression of the outcome on assignment to either treatment, female, and treatment-times-female, controlling for student academic and demographic characteristics and randomization strata dummies (Equation 2). Robust standard errors reported. All beliefs outcomes based on response to post-intervention survey. Realized performance measured at the end of the semester, as percentile rank of final grade.

Table A.11: 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 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.12: 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: 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.13: 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: 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 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.14: Estimated Effect of Intervention on Students' Beliefs, by Pre-Intervention Beliefs

	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
Panel A. Treatment effect by own percentile beliefs									
Students underpredicting	1.562	0.986	2.072	0.006	-0.038	0.043	-0.105***	-0.091*	-0.118***
percentile (got good news)	(1.436)	(2.322)	(1.761)	(0.024)	(0.039)	(0.029)	(0.034)	(0.050)	(0.046)
	[-12.570]	[-11.825]	[-13.252]	[0.144]	[0.219]	[0.081]	[0.585]	[0.500]	[0.658]
Students overpredicting	0.172	0.580	-0.246	-0.047**	-0.038	-0.055**	0.018	0.056	-0.021
percentile (got bad news)	(1.105)	(1.627)	(1.506)	(0.020)	(0.030)	(0.028)	(0.024)	(0.034)	(0.035)
	[13.874]	[15.222]	[12.530]	[0.232]	[0.263]	[0.201]	[0.406]	[0.303]	[0.506]
N	2,032	1,009	1,023	2,223	1,101	1,122	2,223	1,101	1,122
Panel B. Treatment effect by STEM median beliefs									
Students who correctly	1.799	4.472*	-1.175	-0.015	-0.036	0.010	-0.043	0.021	-0.116**
identified STEM median	(1.627)	(2.412)	(2.111)	(0.028)	(0.041)	(0.038)	(0.037)	(0.050)	(0.054)
	[5.075]	[3.764]	[6.544]	[0.173]	[0.213]	[0.129]	[0.407]	[0.335]	[0.486]
Students initially overestimating	0.196	-1.693	1.421	-0.020	-0.018	-0.021	-0.013	0.030	-0.040
median (got good news)	(1.259)	(2.023)	(1.613)	(0.018)	(0.031)	(0.023)	(0.029)	(0.049)	(0.036)
	[6.188]	[11.020]	[3.089]	[0.113]	[0.118]	[0.109]	[0.630]	[0.521]	[0.697]
Students initially underestimating	-0.319	-1.037	0.758	-0.083**	-0.111**	-0.037	-0.039	-0.006	-0.095*
median (got bad news)	(1.950)	(2.523)	(3.081)	(0.038)	(0.049)	(0.061)	(0.032)	(0.040)	(0.052)
	[7.022]	[8.681]	[4.533]	[0.442]	[0.479]	[0.390]	[0.219]	[0.190]	[0.260]
N	2,123	1,036	1,087	2,350	1,142	1,208	2,350	1,142	1,208

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, indicators for pre-intervention beliefs, and treatment-by-pre-beliefs interactions, controlling for student academic and demographic characteristics and randomization strata dummies. Treatment effects by gender estimated from a single regression with a three-way interaction between treatment, female, and pre-intervention beliefs, controlling for student academic and demographic characteristics and randomization strata dummies. Pre-intervention beliefs are based on responses to the pre-intervention survey. In Panel A, underpredicting means the student's self-prediction of their percentile was lower than (or equal to) the percentile the intervention informed them of, while overpredicting means their self-prediction was higher than the information they received. In Panel B, students are categorized by whether they initially correctly identified the course median for students who go on to major in STEM. Robust standard errors reported. Control means in square brackets. All beliefs outcomes based on response to post-intervention survey. Realized performance is measured mid-semester, at the time of intervention.

Table A.15: Estimated Effect of Intervention on Students' Beliefs, by Pre-Intervention Performance

	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
Treatment effect by pre-treatment performance									
Students in bottom quartile	-0.942	-2.598	0.784	-0.055	-0.107**	-0.009	0.033	0.118**	-0.043
at time of treatment	(1.940)	(2.931)	(2.600)	(0.035)	(0.052)	(0.049)	(0.039)	(0.054)	(0.055)
	[21.343]	[25.533]	[16.773]	[0.289]	[0.351]	[0.230]	[0.335]	[0.229]	[0.437]
Students in second quartile	-1.856	-4.593*	0.516	-0.074**	-0.084*	-0.068*	0.029	0.019	0.040
	(1.626)	(2.426)	(2.186)	(0.030)	(0.048)	(0.038)	(0.036)	(0.052)	(0.049)
	[15.167]	[18.900]	[12.111]	[0.236]	[0.295]	[0.189]	[0.406]	[0.327]	[0.469]
Students in third quartile	0.582	2.557	-1.283	-0.012	-0.016	-0.008	-0.069*	-0.041	-0.095*
	(1.536)	(2.139)	(2.184)	(0.029)	(0.044)	(0.037)	(0.038)	(0.054)	(0.054)
	[-0.644]	[0.714]	[-2.070]	[0.169]	[0.202]	[0.135]	[0.533]	[0.465]	[0.604]
Students in top quartile	3.315**	4.858**	1.666	0.003	-0.017	0.024	-0.073**	-0.038	-0.107**
	(1.516)	(2.266)	(1.958)	(0.023)	(0.036)	(0.027)	(0.035)	(0.051)	(0.048)
	[-14.108]	[-13.785]	[-14.459]	[0.112]	[0.163]	[0.062]	[0.597]	[0.473]	[0.721]
N	2,358	1,166	1,192	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, indicators for pre-intervention performance quartile, and treatment-by-performance-quartile interactions, controlling for student academic and demographic characteristics and randomization strata dummies. Treatment effects by gender estimated from a single regression with a three-way interaction between treatment, female, and pre-intervention performance quartile, controlling for student academic and demographic characteristics and randomization strata dummies. Pre-intervention performance is measured at the time of treatment, in November. Robust standard errors reported. Control means in square brackets. All beliefs outcomes based on response to post-intervention survey. Realized performance is measured mid-semester, at the time of intervention.

Table A.16: Estimated Effect of Intervention by Student Level and Intended Major

	Number of STEM credits
A. Treatment effect by student level	
First year or sophomore	-0.211**
	(0.099)
	[8.580]
Junior or senior	-0.051
	(0.269)
	[8.174]
p-value, treat-by-level interaction	0.575
N	5,715
B. Treatment effect by pre-intervention intended major	
Intended STEM major	-0.248**
	(0.123)
	[9.487]
Intended non-STEM major	-0.053
	(0.238)
	[4.809]
p-value, treat-by-major interaction	0.466
N	3,988

Notes: $*p < 0.1$; $**p < 0.05$; $***p < 0.01$. Treatment effects in Panel A estimated from a regression of the outcome on assignment to either treatment, an indicator for whether the student has junior or senior standing, and a treatment-by-level interaction, controlling for student academic and demographic characteristics and randomization strata dummies. Treatment effects in Panel B 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. Student level and course-taking outcomes based on University of Michigan administrative data. Robust standard errors reported. Control means in square brackets.

Table A.17: Estimated Effect of Intervention by Course Subject

	Number of STEM credits
Treatment effect by course subject	
Biology	0.326
	(0.305)
	[7.396]
Chemistry	-0.011
	(0.201)
	[9.534]
Computer Science	-0.431*
	(0.250)
	[8.835]
Economics	-0.165
	(0.255)
	[7.007]
Engineering	0.335
	(0.267)
	[12.763]
Physics	-0.082
	(0.367)
	[12.221]
Statistics	-0.533***
	(0.197)
	[6.771]
P-value, F-test of treat-by-subject interactions	0.080
N	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. Course-taking outcomes based on University of Michigan administrative data.

Table A.18: Estimated Effect of Intervention by Gender Composition of Course (Proportion Men, Continuous)

	Number of STEM credits
Treatment effect	-0.129
(main)	(0.470)
Proportion male	0.499
(main)	(2.973)
Treatment-by-proportion-male interaction	-0.102
	(0.866)
N	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, 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. Robust standard errors reported. Course-taking outcomes based on University of Michigan administrative data.

Appendix B. Randomization-based Inference

In addition to standard inference, I calculate p-values using randomization-based inference. In this approach, randomness in estimates comes from assignment of a fixed number of units (students) to treatment, rather than from sampling from a population.

To implement, I re-assign treatment status 10,000 times, using the same procedure used in the original randomization. This accounts for the fact that my re-randomization procedure changes the distribution of test statistics, because I discard any re-randomizations that do not meet the pre-specified balance rule (Bruhn and McKenzie, 2009). Randomization inference also addresses concerns about clustered data, because it preserves the underlying data structure, including any mean or higher-order correlations. Under each "treatment" assignment, I calculate a test statistic of interest (a main effect, the effect for men, the effect for women, or the differential effect). This process generates a distribution of potential treatment effects that could be due to baseline differences between students assigned to treatment and control. (Note that this accounts for any outliers that may be driving treatment effects.) For each effect, I calculate the share of the 10,000 simulated estimates that are larger in absolute value than the estimate observed under the true treatment assignment; this proportion represents the randomization-based p-value. Note that while the traditional sampling approach tests a null hypothesis of no average effect, randomization inference tests a sharp null hypothesis of no effect for any individual.

A comparison of sampling or model-based p-values and randomization-based p-values is presented in Appendix Table B.1. Although they represent different conceptual approaches, the model- and randomization-based p-values produce virtually identical conclusions.

Table B.1: 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
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
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

Notes: Each pair of p-values correspond to a single test statistic. Model-based p-values correspond to the analyses in Tables 2 and 3. 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 C.1: 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
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
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

Notes: Each row corresponds to a single test statistic. Effects and unadjusted p-values correspond to the analyses in Tables 3 and 2. 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).