Economics of Education Review 97 (2023) 102479

Contents lists available at ScienceDirect

Economics of Education Review

journal homepage: www.elsevier.com/locate/econedurev

College major choice and beliefs about relative performance: An experimental intervention to understand gender gaps in STEM

Stephanie Owen *

Colby College Department of Economics, Waterville, ME, USA

ARTICLE INFO

Dataset link: https://enrollment.umich.edu/data/learning-analytics-data-architecture-larc

JEL classification:
I21
I24
J16
D91

Keywords:
STEM
Beliefs
Gender
College major

ABSTRACT

Beliefs about relative academic performance may shape college major choice and explain gender gaps in STEM, but little causal evidence exists. To test whether these beliefs are malleable and salient enough to change behavior, I run a randomized experiment with 5,700 undergraduates across seven introductory STEM courses. Providing relative performance information shrinks gender gaps in biased beliefs substantially. However, students' course-taking and major choice are largely unchanged. If anything, initially overconfident men and women were discouraged by the intervention. Increasing female STEM participation may require more intensive or targeted intervention.

1. Introduction

Understanding how individuals make decisions about college major 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 & Riordan, 2012), and research has documented shortages in certain skills and sectors (Xue & Larson, 2015). In addition to overall shortages, women remain persistently underrepresented in many quantitative fields such as economics, engineering, and computer science. Although they represent more than half of all college graduates, women receive only a third of bachelor's degrees in economics and approximately a fifth of degrees in engineering and computer science (author's calculations using 2017 IPEDS data).

The gender gap in STEM education has implications for both equity and efficiency. The fields with the fewest women also tend to be the highest-paying ones, so differences in 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 (Ceci et al., 2014; Cheryan et al., 2017), 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 (Beyer, 1990; Beyer & Bowden, 1997; Exley & Kessler, 2022; Lundeberg et al., 1994; Marshman et al., 2018; Niederle & Vesterlund, 2007; Page & Ruebeck, 2022; Vincent-Ruz et al., 2018), 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 (Bobba & Frisancho, 2019; Franco, 2019; Gonzalez, 2017; Hakimov et al., 2022; Wozniak et al., 2014). Several recent field experiments have shown that it is possible to change the academic decisions of college students and close gaps in major choice with light-touch interventions, though cannot fully disentangle the mechanisms responsible or the reasons for gender differences (Bayer et al., 2019; Li, 2018; Porter & Serra, 2019). Together,

* Correspondence to: Department of Economics, 5237 Mayflower Hill, Waterville, ME 04901, USA.
E-mail address: sowen@colby.edu.

https://doi.org/10.1016/j.econedurev.2023.102479
Received 26 May 2023; Received in revised form 18 September 2023; Accepted 26 September 2023
Available online 12 October 2023

these prior strands of work suggest that beliefs about performance may be malleable and salient enough to affect the college major choices of underrepresented groups, but causal evidence on this mechanism has thus far been limited.

This paper provides large-scale experimental evidence isolating the effect of beliefs about relative performance on college major choice, 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. (Throughout the paper, references to STEM include economics.) 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. I combine these survey data with administrative data on students' course-taking and major choice.

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.

I also find striking and persistent gender differences in students' accuracy in identifying the median course grade for students who go on to major in STEM. Men are about ten percentage points more likely to think the median course grade for students who go on to major in STEM is lower than it actually is, while women are about 20 percentage points more likely to think it is higher than it is. The patterns in this second type of belief, which no other study has measured, imply male overconfidence and female underconfidence about their performance relative to others. A correlational exercise with students in the control group indicates that these two types of beliefs may account for approximately seven percent of the two-credit (half of a course) gender gap in STEM course-taking in the subsequent semester and 15 percent of the gap in major choice, even controlling for realized performance and a rich set of academic and demographic characteristics. In this exercise, beliefs explain as much (or more) of the gap as does prior math achievement.

Providing information on actual relative performance causes 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. 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.

I find limited evidence that the informational intervention changed shorter or longer term behavior. In the semester following the intervention, men decreased the number of STEM credits they took by 0.3 (three percent). However, I detect no change in the subsequent four semesters. I also detect little change in women's STEM course-taking; over five semesters, effect sizes are null with the exception of a 0.4 credit (six percent) decrease in the fourth post-intervention semester. In the five semesters following the intervention, I find no change to the probability that either men or women declare a STEM major. All of the point estimates on major choice are negative but statistically insignificant, suggesting a possible small discouragement effect. Heterogeneity by pre-intervention beliefs suggests that students who received bad news about their relative performance – both men and women – were discouraged by the information. The modest changes to behavior do not appear to be driven by changes in students' class performance, stress about grades, or STEM self-efficacy beliefs.

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 or major choice behavior. For this reason, the majority of the results I present combine the two treatment arms and reflect a general effect of information provision.

This study provides, to my knowledge, the largest scale evidence on the causal effect of beliefs and belief updating on college major choice and the gender gap therein. Existing evidence has thus far been limited by small sample size, narrowness of the population studied, and a lack of real world, long-term follow-up data. The combination of a large-scale field experiment, a setting covering multiple STEM disciplines, survey data on beliefs, and long-term administrative follow-up data represent a significant contribution to this much-studied topic.

As a whole, my experimental results suggest that while stark gender differences in beliefs exist, and it is possible to debias them, light-touch information provision has a limited effect on behavior and the male–female STEM gap. If anything, information may discourage overconfident students of all genders. However, I cannot rule out that a more targeted or more intensive intervention could have larger, more positive effects. It is also possible that a similar intervention with students who are younger than college-age – and therefore might have more malleable beliefs and behavior – could be more effective.

The paper proceeds as follows. I summarize related literature in Section 2, introduce the setting and data in Section 3, describe the experiment in Section 4, and lay out empirical methods in Section 5. I present my results in Section 6. Section 7 contextualizes the results and Section 8 concludes.

2. Related literature

Understanding and closing the gender gap in major choice has been the focus of much speculation and research (see Delaney and Devereux (2021) for a review). Candidate mechanisms that may explain differential rates of participation and persistence in STEM include: mathematical aptitude and comparative advantage (Aucejo & James, 2021; Breda & Napp, 2019; Speer, 2023), risk aversion and willingness to compete (Buser et al., 2014, 2017; Niederle & Vesterlund, 2007), (lack of) female role models (Bettinger & Long, 2005; Carrell et al., 2010), gender composition of peers (Booth et al., 2018; Bostwick & Weinberg, 2022), interest and relevance of the topics/curriculum (Jensen & Owen, 2000; Owen & Hagstrom, 2021), preference for certain types of jobs and job characteristics (Kuhn & Wolter, 2022; Wiswall & Zafar, 2015; Zafar, 2013), discrimination and bias (Avitzour et al., 2020; Carlana, 2019), toxic culture and harassment (Aycock et al., 2019; Minnotte & Pedersen, 2023), response to grades and academic feedback (Avilova & Goldin, 2020; Kugler et al., 2021; Owen, 2010), and method of assessment (Azmat et al., 2016; Griselda, 2022; Iriberri & Rey-Biel, 2021). A decision as consequential and complex as field of study is almost certainly determined by many factors at multiple points in time. Rigorous causal evidence attempting to isolate many of the above mechanisms has been limited and come to mixed conclusions (Delaney

& Devereux, 2021). Given the scope of the problem, there is particular interest in policies that could increase female STEM participation at scale. Informational interventions are especially appealing given their low-cost, easily scalable nature.

Several strands of research point to confidence and beliefs about performance and ability as a key driver of gender differences in major choice and a promising target of intervention. A seminal study by Niederle and Vesterlund (2007) found that men were more overconfident than women about their relative performance on a number-adding task, and this explained much of the gender gap in competitiveness, with men choosing to enter a tournament style of compensation much more than women; the authors hypothesize that this pattern could apply to the choice to enter competitive fields such as engineering. Lab studies in psychology have documented similar gaps in confidence, with male participants reporting overly positive self-evaluations on a male-typed task and female participants reporting overly negative self-evaluations (Beyer, 1990; Beyer & Bowden, 1997). More recent work found that among students in introductory science courses, male and female students with similar levels of knowledge and performance report different levels of self-efficacy.[1] Marshman et al. (2018) found that women with A's in introductory physics courses had self-efficacy levels on par with those of men earning C's. A study with introductory chemistry students similarly found that among students with similar SAT math scores, men had higher confidence in their ability to be successful in chemistry (Vincent-Ruz et al., 2018). Exley and Kessler (2022) find, in an online lab setting, that women describe their performance on a math and science task less positively than equally performing men, and that this leads to (simulated) employment and earnings gaps by gender. In a second analysis, they survey middle and high school students and find that gender differences in self-evaluation appear as early as sixth grade.

Although the gender differences in these studies are striking, none has connected the studied beliefs with real-world decisions such as STEM persistence or major choice. One exception is Page and Ruebeck (2022), who use PSID survey data and find a link between a childhood (age 8–11) measure of confidence in math ability and adult outcomes including college major and earnings. Although the confidence measure exhibits significant gender differences, the authors cannot make causal claims, and the data used lack the precision needed to conclude that math confidence explains gender gaps in later outcomes.

Although empirical causal evidence is limited, two canonical economic frameworks provide strong theoretical motivation for the importance of performance beliefs in college major decisions. 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., Coffman et al., 2019; Mobius et al., 2014). 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.

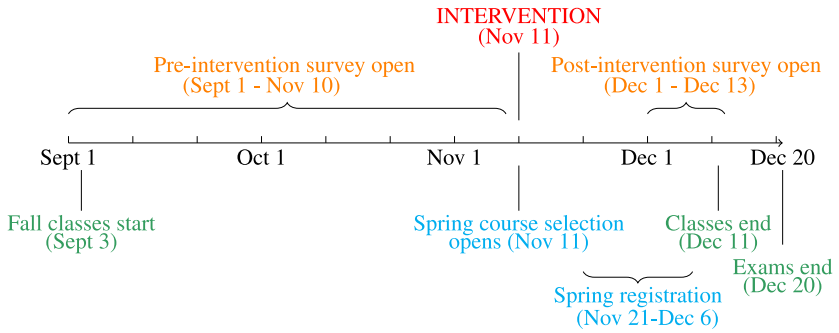
An implication of these models is that, assuming a positive relationship between beliefs about major-specific ability and the expected payoff to a major, those performing better in STEM than they expected should be (weakly) more likely to pursue STEM, while those who receive a negative signal should 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.

Evidence from other settings has shown that it is possible to de-bias beliefs and change behavior by providing more accurate information about performance. A lab experiment by Wozniak et al. (2014) provided relative performance feedback to participants and closed the gender gap in the choice to compete; high-ability women increased and low-ability men decreased their likelihood of entering a tournament form of compensation. In a field experiment, Hakimov et al. (2022) told French high school students applying to college their rank in the national grade distribution, thus closing gender gaps in application behavior. A similar experiment by Bobba and Frisancho (2019) in Mexico City found that providing information on absolute and relative performance on a high school admissions test led students to have more accurate beliefs and update their choice of high school track to be more in line with ability. Gonzalez (2017) found that high school students revised their academic plans to take advanced coursework in response to being told they had the potential to do well in Advanced Placement courses. Franco (2019) provided performance feedback to Colombian students prepping for a college entrance exam; low-performing students appeared to be discouraged by the information, reducing effort and the choice to take the exam. These studies imply that it is possible to change educational choices by moving beliefs about performance and ability, but none have shown this in the context of college students choosing a major.

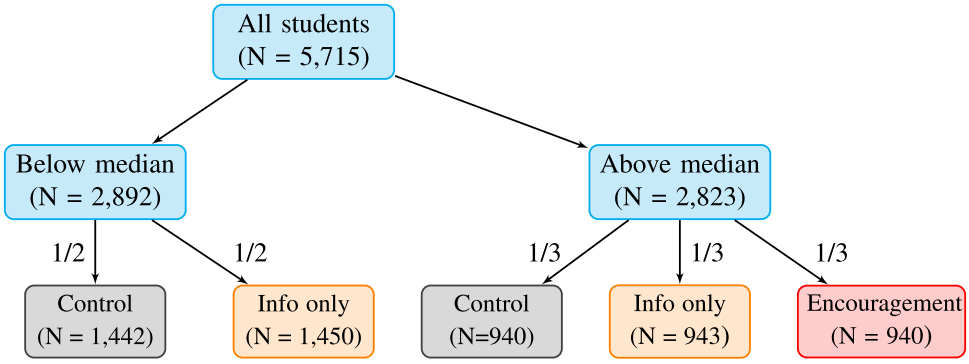
Finally, there is a small but growing literature of interventions aiming to attract and retain students from underrepresented groups in undergraduate economics programs. Li (2018) implemented an intervention among introductory economics students that bundled several mechanisms: information about relative performance, encouragement to major in economics, and information about the field of economics. As a result, high-performing female students were more likely to major in economics, and low-performing male students were less likely. However, the treatment arms varied by student gender and performance. The experimental design of Li (2018) is such that it cannot separately identify the effects of performance information versus 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) invited recent alumnae to visit an undergraduate economics class to talk about their current jobs and the role economics played in their careers. It had a large effect on female students' likelihood of taking further courses and majoring in economics. 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) sent incoming college students a welcoming email and information about the field of economics, which increased the likelihood that an underrepresented student enrolled in an economics course. However, they only targeted women, first-generation students, and underrepresented minorities, so cannot say whether White and Asian men would react similarly. These interventions have largely been limited to economics courses and students rather than a broader set of male-dominated subjects, and there is more work to be done on precise mechanisms, but they prove that fairly light-touch intervention can successfully affect major choice.

Taken as a whole, all of the above research provides support for a promising but largely untested avenue of intervention. The theoretical and empirical evidence suggests that it is possible to update students' beliefs by providing information about their relative performance, and that doing so could alter their choice of academic major in a way that shrinks gender gaps in STEM. However, there is as of yet no causal evidence isolating this mechanism in a real world college setting. Furthermore, existing evidence tends to come from small samples, within a single field (e.g., economics), and have limited follow-up

[1] In the educational psychology literature, self-efficacy is defined as "the belief in one's capability to be successful in a particular task, course, or subject area" Marshman et al., 2018, p. 020123–1.



(a) Study Timeline



(b) Experimental Design

Fig. 1. Study timeline and 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.

data.[2] The current study marries these literatures and provides the first and largest scale causal evidence across multiple STEM fields and with several years of follow-up data.

3. 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.[3] The courses span seven departments and subjects: biology, chemistry, computer science, economics, engineering, physics, and statistics.[4]

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

I use two main sources of data. The first is student administrative records from UM (University of Michigan Office of Enrollment Management, 2022). 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

[2] For example, some of the most well-cited studies are Wiswall and Zafar (2015), which has a sample size of fewer than 500; Zafar (2013), which has a sample size 161; and Niederle and Vesterlund (2007), which has a sample size of only 80. The previous three studies were conducted in a lab setting, so can only speculate about effects on real-world decisions. The field experiment that is arguably closest to the current study is Li (2018), with N=450. The field interventions by Li (2018), Porter and Serra (2019), and Bayer et al. (2019) all focus on economics students only. Li (2018) and Bayer et al. (2019) each only report one year of follow-up data.

[3] A second round of the study, planned for the spring semester of 2020, was canceled due to the COVID-19 pandemic.

[4] 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).

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 (standardized)	−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 high school	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	
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 are based on a regression of the characteristic on treatment status, controlling for randomization strata. I also test for differences in missingness rates on all variables and find none. The 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 are based on University of Michigan administrative data.

the post-intervention survey in December.[5] 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.[6] Three courses (biology, computer science, and engineering) offered credit for the post-intervention survey.

4. Experimental design

4.1. Intervention

The intervention consisted of two treatment arms, which I refer to as information-only and information- plus-encouragement. The two treatment arms were delivered as online messages and emails to students.[7] 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. (For a detailed timeline, see panel (a) of Fig. 1.)

The first treatment arm, the information-only intervention, provided 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,[8] and (again) students who go on to

[5] The pre-intervention survey remained open to students throughout the semester, but I drop any responses from after the intervention. See Fig. 1, panel (a) for a timeline of data collection.

[6] 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.

[7] This study, including an analysis plan, 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.

[8] 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.

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.[9] Appendix Figure A.2 shows an example of an information-plus-encouragement message. In designing a second treatment arm, I wanted to test whether the framing of the information affected how students incorporated it. The findings of Li (2018), an experimental intervention that bundled relative performance information with encouragement and information about the field of economics, suggest that the encouragement aspect may be important for high-performing women in particular but cannot disentangle the various components.[10]

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.

4.2. Treatment assignment and take-up

I assigned treatment status at the student level, stratified by course, gender, and performance at the time of randomization (above versus below the course median). This results in $8 \times 2 \times 2 = 32$ strata.[11] Within each of the 16 below-median strata, the probability of receiving the information-only treatment was 0.5. Students who were above the median were eligible for the second treatment arm; within the 16 above-median strata, the information-only and information-plus-encouragement treatment were 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-randomized until each pre-treatment characteristic was balanced within strata (minimum p-value of 0.1). 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. Fig. 1, panel (b) summarizes the experimental design.[12]

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 delivery formats—including the visual element, the histogram—was identical.

Among students who were sent a treatment message, 83 percent viewed it in some format. 57.5 percent viewed the message only as an email, three percent saw the message only within ECoach, and 23 percent viewed it in both formats.[13] 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. Note that because opening or scrolling through a message does not necessarily indicate a close read of the content, I consider these view rates to be upper bounds for "receiving" the information.[14]

4.3. 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.[15]

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). The sample demographics largely reflect the demographics of the university, though male, White, and Asian students are 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).[16] The average UM student and the average student in this sample come from a socioeconomically advantaged background: 60.5 percent have a parent with a graduate or professional degree, and only 15 percent are first-generation (meaning neither parent has a bachelor's degree). The majority (64 percent) have family incomes above $100,000. Roughly half of the sample (52 percent) are Michigan residents.

The average cumulative GPA while at UM is 3.41 (students in their first semester do not yet have values for this variable). UM is a highly selective school, and this is reflected in the high average test scores

[9] 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.

[10] Li (2018)'s intervention had a positive effect on high-performing women, who received relative performance information, encouragement to major in economics, and information about the field of economics. 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.

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

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

[13] I was able to track email views via a hidden pixel in the intervention message, and ECoach views via site metadata.

[14] 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.1). 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.

[15] Table 1 pools students receiving either treatment; a balance table that separates the two treatment arms is presented in Appendix Table A.2. I also test for balance separately by gender in Appendix Table A.3.

[16] 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.

Table 2
Decomposition of gender gap in STEM credits and major choice by relative performance beliefs and other covariate components (control students only).

	STEM credits (1 semester later)		ℙ(STEM major) (5 semesters later)	
Raw female–male gap	−2.148		−0.142	
	(0.280)		(0.031)	
Covariate	Gap explained by covariate	Percent of total gap	Gap explained by covariate	Percent of total gap
Own percentile belief	−0.044	2%	−0.003	2%
	(0.037)		(0.004)	
STEM median belief	−0.105	5%	−0.019	13%
	(0.052)		(0.007)	
Realized percentile	−0.021	1%	−0.002	2%
	(0.024)		(0.003)	
Demographics	0.017	−1%	0.001	−1%
	(0.050)		(0.006)	
High school achievement	−0.021	1%	−0.011	8%
	(0.097)		(0.011)	
Math placement score	−0.146	7%	−0.016	11%
	(0.059)		(0.007)	
Prior college achievement	−0.039	2%	−0.007	5%
	(0.047)		(0.006)	
Student level	0.002	0%	0.001	0%
	(0.025)		(0.004)	
Declared major	−0.690	32%	−0.032	23%
	(0.155)		(0.014)	
Total explained	−1.048	49%	−0.088	62%
Total unexplained	−1.100	51%	−0.054	38%
N	918		918	

Notes: This decomposition follows Gelbach (2016). STEM credits are measured in the semester following the one when students took the course. STEM major is measured five semesters later; a student who is declared as a STEM major or graduated with a STEM degree is considered a major. Own percentile belief is a student's 1-100 prediction of their own final course percentile, measured in the end of semester survey. STEM median belief is measured as two indicators for whether a student is over- or underestimating the course median for STEM majors, measured in the end of semester survey. Demographics include race, parent education, family income, and in-state status. High school achievement includes ACT and SAT scores, high school GPA, and a high school calculus indicator. College achievement is measured as prior GPA at UM. The sample is limited to control students who answered both surveys.

(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.[17] Based on beginning-of-semester survey data (not shown), 72 percent of women and 85 percent of men enter these courses intending to pursue a STEM major.

I test for treatment-control 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.

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.4.

4.4. Survey response

Around three quarters of students responded to the pre-intervention survey, while slightly fewer than half (48.7 percent) responded to the post-intervention survey.[18] Crucially, survey response does not vary by treatment status. While there is non-random selection into survey response, the selection is similar in the treated and control groups.

I 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). I focus on the post-intervention survey here, since I estimate treatment effects on post-intervention variables. 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 (recall that instructors in these courses offered extra credit for both surveys), 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 and major choice outcomes, which use administrative data. However, survey non-response 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 4.3, 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 re-estimate effects on survey belief outcomes using inverse probability weighting to make survey respondents resemble the full sample on their observable characteristics. In the second, I estimate treatment effects on administrative data outcomes using only the sample who responded to the survey. In both cases, I lose precision but the point estimates are similar.

5. Empirical method

For my experimental analysis, I estimate the main effect of the intervention with the following specification:

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

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 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:

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

[17] 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.

[18] 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 percent (for intended major).

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	0.460	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 are estimated from a regression of the outcome on assignment to either treatment, controlling for student academic and demographic characteristics and randomization strata dummies (Eq. (1)). Treatment effects by gender are 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 (Eq. (2)). Robust standard errors are reported. All beliefs outcomes are based on responses to the post-intervention survey. Realized performance is measured mid-semester, at the time of intervention.

Table 4
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 percentile (got good news)	1.562	0.986	2.072	0.006	−0.038	0.043	−0.105***	−0.091*	−0.118***
	(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 percentile (got bad news)	0.172	0.580	−0.246	−0.047**	−0.038	−0.055**	0.018	0.056	−0.021
	(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 identified STEM median	1.799	4.472*	−1.175	−0.015	−0.036	0.010	−0.043	0.021	−0.116**
	(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 median (got good news)	0.196	−1.693	1.421	−0.020	−0.018	−0.021	−0.013	0.030	−0.040
	(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 median (got bad news)	−0.319	−1.037	0.758	−0.083**	−0.111**	−0.037	−0.039	−0.006	−0.095*
	(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 are 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 are 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 are reported. Control means are in square brackets. All beliefs outcomes are based on responses to the post-intervention survey. Realized performance is measured mid-semester, at the time of intervention.

Table 5
Estimated effect of intervention on students' STEM major choice, by pre-intervention beliefs.

	Declared as STEM major five semesters post intervention		
	All	Men	Women
Panel A. Treatment effect by own percentile beliefs			
Students underpredicting	0.027	0.032	0.024
percentile (got good news)	(0.024)	(0.034)	(0.035)
	[0.660]	[0.706]	[0.619]
Students overpredicting	−0.036**	−0.025	−0.049*
percentile (got bad news)	(0.017)	(0.023)	(0.026)
	[0.603]	[0.702]	[0.494]
N	3,664	1,874	1,790
Panel B. Treatment effect by STEM median beliefs			
Students who correctly	−0.022	−0.004	−0.044
identified STEM median	(0.025)	(0.032)	(0.038)
	[0.645]	[0.733]	[0.545]
Students initially overestimating	−0.006	0.020	−0.025
median (got good news)	(0.020)	(0.030)	(0.027)
	[0.568]	[0.660]	[0.501]
Students initially underestimating	−0.033	−0.054*	0.003
median (got bad news)	(0.027)	(0.032)	(0.046)
	[0.698]	[0.741]	[0.626]
N	3,915	1,973	1,942

Notes: $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$. Treatment effects for all students are 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 are 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 are reported. Control means are in square brackets. Major declaration outcomes are based on University of Michigan administrative data.

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

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

$$Y_i = \beta_0 + \beta_1 Info_i + \beta_2 Encourage_i + \boldsymbol{\gamma} \boldsymbol{X}_i' + \delta_s + \epsilon_i \quad (3)$$

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 Eq. (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 effects on students' beliefs about their relative performance using outcomes measured in the post-intervention survey. I estimate treatment effects on course-taking (number of STEM credits) and major choice (declaration of a STEM major) based on administrative transcript data. I investigate additional mechanisms using outcomes and characteristics collected in the survey and available in administrative data. All results report robust standard errors and significance levels.

5.1. Outcome measures

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.[19] 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 in a STEM field; I can then compare their answers to the true median.[20] 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.

My primary behavioral outcomes are course-taking, operationalized as the number of STEM credits attempted, and STEM major declaration.

[19] 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.

[20] 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.

Table 6
Estimated effect of intervention on performance and academic attitudes.

	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 (standard deviation units)			Grade stress (standard deviation 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	0.000	0.116	−0.108	0.000	−0.239	0.221
N	2,687	1,317	1,370	2,638	1,290	1,348
	Intent to major in STEM (binary)			STEM interest/intent index (standard deviation 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	0.000	0.110	−0.102
N	2,662	1,302	1,360	2,639	1,289	1,350

Notes: $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$. Treatment effects for all students are estimated from a regression of the outcome on assignment to either treatment, controlling for student academic and demographic characteristics and randomization strata dummies (Eq. (1)). Treatment effects by gender are 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 (Eq. (2)). Robust standard errors are reported. Performance outcomes are based on University of Michigan administrative data. The 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. STEM interest and intent outcomes are based on responses to the post-intervention survey.

I have five semesters of follow-up data (through the spring 2022 term) and show treatment effects by semester. 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.[21] These outcomes come from the administrative data; attrition or missingness occurs only if a student graduates or drops out. If a student graduates with a degree in a STEM field, they are coded as a declared STEM major for all subsequent semesters.[22]

6. Results

6.1. Control students' beliefs about relative performance

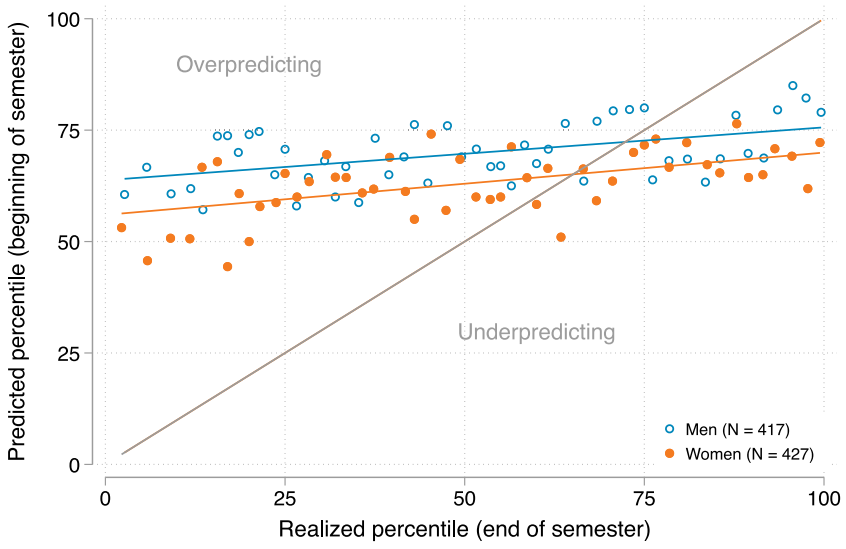
To motivate the experimental results, I begin by describing students' beliefs in the absence of any intervention. In this section, I focus on control students only. I examine control students' beliefs at two points in time: at the beginning of the semester (generally in September) and again at the end 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.[23] 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 (all reported differences are statistically significant). Low-performing (below-median) students tend to overestimate their performance (by 30.6 percentiles), while high-performing ones

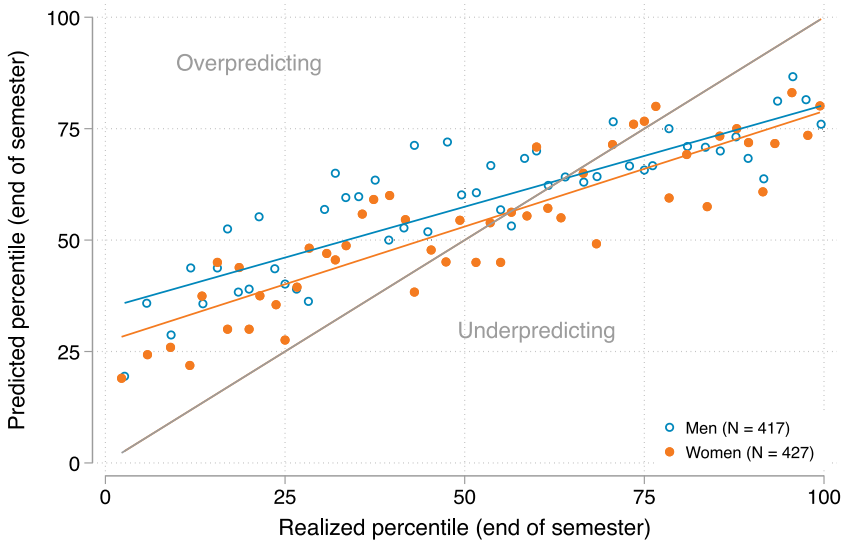
[21] I code the following subjects (CIP codes) as STEM: natural resources and conservation (03), computer and information sciences (11), engineering (14), biological and biomedical sciences (26), mathematics and statistics (27), physical sciences (40), and economics (45.06). I code economics (45.06)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

[22] 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.

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



(a) Beginning of Semester Beliefs



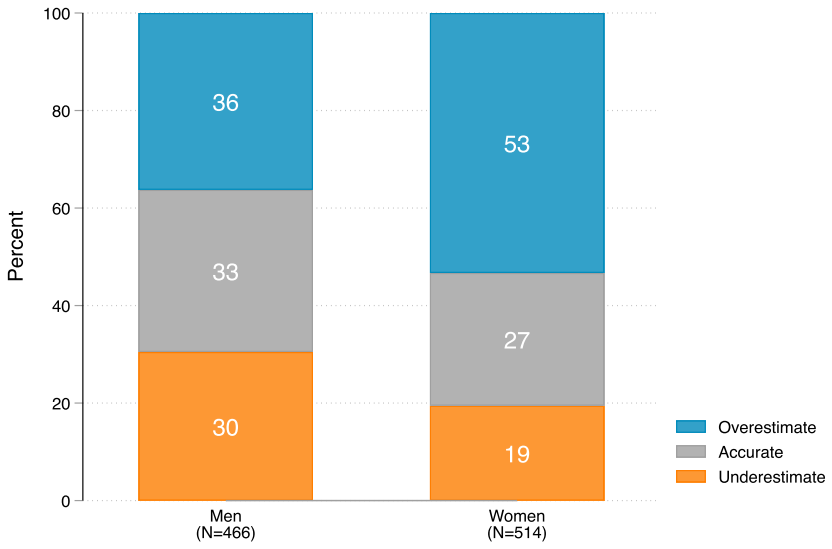
(b) End of Semester Beliefs

Fig. 2. Control student beliefs about own percentile, by gender.
Notes: The sample is 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. Both figures are binned scatterplots, plotting average predicted percentile within 50 equally-sized bins of students, grouped by realized percentile.

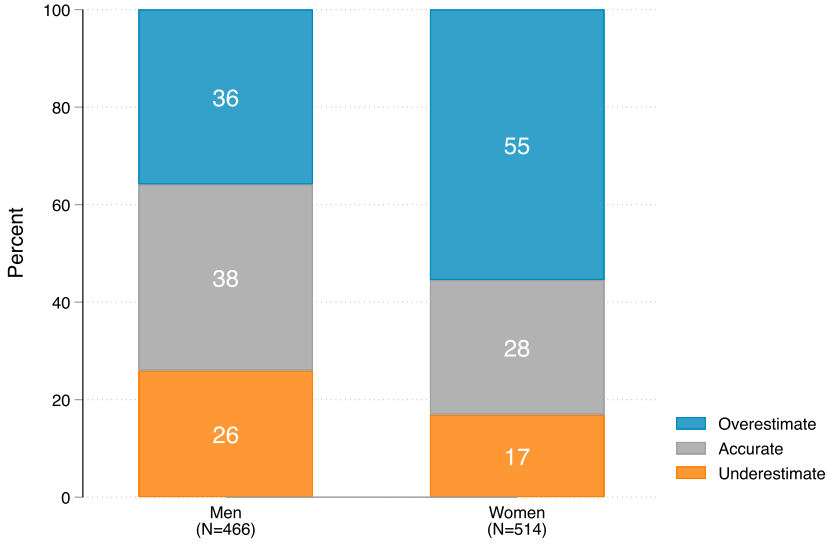
tend to underestimate, though to a lesser extent (average underprediction of 2.7 points).[24] Low-performing men are the most overconfident (overpredicting by an average of 34.4 percentiles, compared to 27.2 for low-performing women) while high-performing women are the most underconfident (underpredicting by 5.9 percentiles compared to less than a percentile for high-performing men). Panel (a) of Fig. 2 visually 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 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

[24] 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.



(a) Beginning of Semester Beliefs



(b) End of Semester Beliefs

Fig. 3. Control student beliefs about course median for STEM majors, by gender.
Notes: The sample is restricted to control students who responded to the question about the median on both the pre- and post-intervention surveys. The median refers to the median grade for students who previously took the course and later majored in a STEM field. The median is not gender-specific. Overestimating means the student thinks the median is higher than it is (e.g., the median is a B and they think it is a B+), while underestimating means they think the median is lower than it is.

not statistically significant. These changes are reflected in Panel (b) of Fig. 2.

I next turn to what students believe about the performance of STEM majors. Panel (a) of Fig. 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; overestimating means the student thinks others are doing better than they 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 (Fig. 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).[25]

[25] 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,

The two sets of findings about control students' beliefs – about their own relative rank and about the performance of other STEM majors – work in the same direction, and support 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.) By five semesters later, men are 14 percentage points more likely to be declared as STEM majors.

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 investigate more systematically whether beliefs about relative performance are related to the gender gap in course-taking and major choice, I perform a decomposition following Gelbach (2016). This accounting exercise uses the omitted variable bias formula to partial out how much the addition of a variable to a regression changes some base coefficient—in this case, the coefficient on female, which represents the gender gap.[26]

I apply the decomposition to a model where I regress STEM persistence outcomes on a female indicator, demographic and academic controls (those listed in Table 1), the student's final percentile rank in the course, their prediction of their final percentile, and indicators for whether they are under- or overestimating the median course grade for STEM majors. I examine the gaps in two outcomes: number of STEM credits in the semester following the course, and the likelihood of being declared as a STEM major five semesters later. Although this approach shows the relationship between beliefs and behavior after controlling for a number of potentially confounding factors, I do not assign a causal interpretation to these results; rather, I use them as motivating evidence for my experimental analysis. Only control students who responded to both surveys are included in this exercise.

The results, in Table 2, show that the full set of belief, performance, academic, and demographic variables account for roughly half of the observed gender gap in credits (2.15 credits in this sample) and more than 60 percent of the 14-point gap in major choice. Students' beliefs about their own course percentile explain around two percent of the gender gap in credits, and beliefs about the course median for STEM majors explain an additional 5 percent. Together, the beliefs measures account for seven percent of the total gender gap—the same amount explained by their college math placement score. The decomposition of the gap in major choice produces even stronger conclusions, with the two types of beliefs statistically explaining 15 percent of the gap.

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

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

6.2. *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 6.1. 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. I test for effects on performance directly in Section 6.5.[27] I report both an absolute value measure as well as a signed measure that captures the direction of the error. Second, I measure the accuracy of beliefs about the performance of STEM majors by creating two indicator variables for whether a student is over- or underestimating the course median for students who go on to major in STEM.

Table 3 shows treatment effects on beliefs outcomes, for the full sample as well as separately for men and women.[28] 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 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 error 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 Fig. 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. I test for heterogeneity in beliefs more formally in Section 6.4.

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

but the differences by gender are much smaller. Among control men, 55 percent underestimate, 33 percent are accurate, and 12 percent overestimate the overall median at the end of the semester. Among control women, the proportions are 50, 35, and 15 percent. The negligible gender differences in overall median beliefs imply that it is not the case that men and women have different beliefs about grades or grade inflation generally. Rather, they hold different beliefs about the selection into STEM, with women setting the bar for STEM higher.

[26] An advantage of this approach relative to one that progressively adds covariates is that it is not sensitive to the order in which covariates are added. The Gelbach decomposition is conceptually similar to a Kitagawa-Oaxaca-Blinder decomposition, and in fact is equivalent once interactions between the covariates and gender are included.

[27] 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.8). The signs are similar but the magnitudes somewhat smaller. This is not surprising given that the intervention told students their mid-semester percentile; they updated their beliefs in the direction of the signal they received.

[28] Treatment effects on beliefs outcomes estimated without covariates are included as Appendix Table A.9. The results are very similar.

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, though the gender differences in treatment effects do not reach conventional levels of statistical significance.[29]

6.3. Effect of intervention on STEM persistence

Fig. 5 summarizes the effect of the intervention on students' course-taking (number of STEM credits, shown in panel (a)) and major choice (panel (b)), with treatment effects estimated for each available post-treatment semester. (A table of estimated effects, standard errors, and control means appears as Appendix Table A.11.) I find a small negative effect (−0.28 credits or three percent) of the intervention on men's course-taking in the semester immediately following the intervention. However, the effect disappears in later semesters, and none are statistically distinguishable from the effects for women. I also estimate a negative effect (−0.39 credits or six percent) for women in the fourth post-intervention semester. I find little evidence that either men or women changed their choice of major; although all point estimates are negative (except for one estimate that is zero to three decimal places), none are statistically significant.[30] However, the results are somewhat imprecise. The 95 percent confidence intervals on major choice five semesters post-intervention imply that I can't rule out positive effects as large as 1.9 percentage points for women (3.6 percent relative to the control mean of 51.8 percent) or negative effects as large as −4.7 (−9 percent). For men, the confidence interval ranges from −3.9 to 2 percentage points (−5.6 to 2.9 percent, off a control mean of 69.5 percent).

For high-performing students, who were eligible for the second treatment arm, I test for differential effects on STEM course-taking and major choice by treatment arm (Appendix Table A.12) but find none, for women or men.[31] 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.

6.4. Heterogeneity by pre-intervention beliefs

We might expect that students who were initially overconfident about their relative performance and for whom the informational intervention contained bad news to react differently than those who received good news. To better understand how the intervention caused updating to beliefs and behavior, I estimate heterogeneous treatment effects based on students' initial beliefs in the pre-intervention survey. 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 (for most students) 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. Second, 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. All of these results rely on survey data with considerable missingness, so should be interpreted with some caution.

Table 4 shows differences in belief updating for students with different initial beliefs. Panel A shows how belief updating differs by initial under- vs. over-prediction of their own percentile. The patterns in own-percentile belief updating do not show the expected pattern. Both those who got good news as well as those who received bad news updated their percentile beliefs slightly upwards, and none of the effects (or differences between effects) are statistically significant. Students who were initially underconfident about their percentile adjusted their belief about the course median downward (10.5 percentage points less likely to overestimate); though these are different measures, the pattern is consistent in that underconfident students (in terms of percentile beliefs) updated in a way that corrected underconfidence (in terms of median beliefs). Similarly, students who were initially overpredicting their percentile and received bad news corrected their belief about the median in a way that corrected overconfidence, becoming 4.7 percentage points less likely to underestimate the median. I cannot reject equality of effect by gender for any of the effects.

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. 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. For example, initially accurate and underestimating women became less likely to overestimate the median. I also lack the statistical power to say whether men and women update differently.

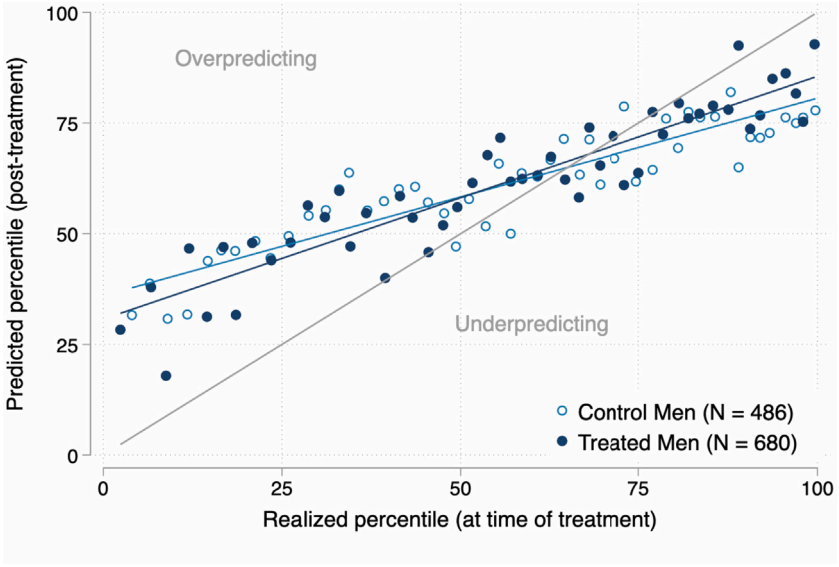
In Table 5, I examine the same heterogeneity but with major declaration five semesters post-intervention as the outcome. These results are suggestive that the initially overconfident students, who received bad news about their relative performance, may have reacted by switching out of a STEM major. By five semesters later, treated students who had initially overpredicted their percentile were 3.6 percentage points less likely to be declared as a STEM major than equivalent control students. The point estimate is larger for women (−4.9 vs. −2.5 percentage points), but I can't reject that it's statistically equivalent. I also find weak evidence that men who got bad news about the course median for STEM majors were less likely (by 5.4 percentage points) to be a STEM major as of the last follow-up, though again I cannot reject the possibility that women responded equally.

[29] As a 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.10. The point estimates are similar to the ones in Table 3, though somewhat less precise.

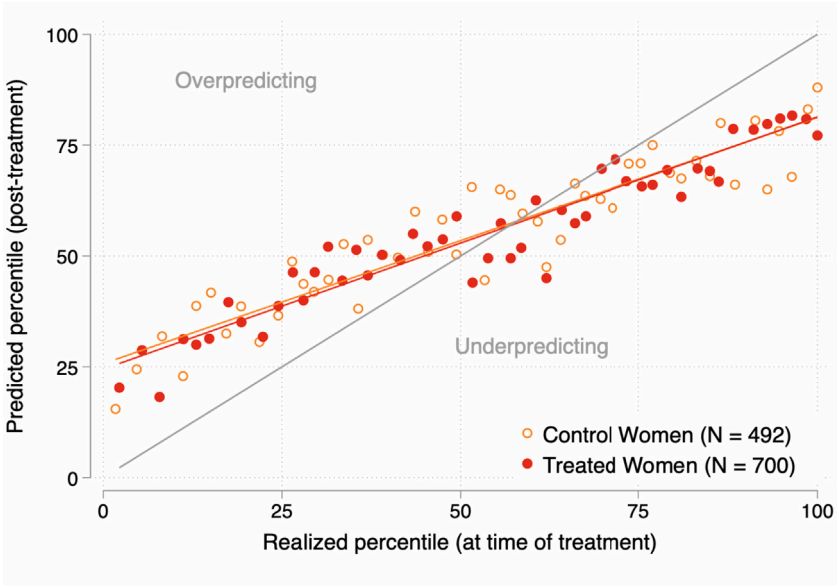
[30] Treatment effects on STEM course-taking and major choice estimated without covariates are included as Appendix Figure A.4. The results are very similar. I also include treatment effects estimated using only students who responded to the post-intervention survey, in Figure A.5. Again, the results are similar. This exercise, along with Appendix Table A.10, suggests that differential survey response is not leading to a spurious conclusion about the relationship between changes to beliefs and changes to behavior.

[31] 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 power to distinguish between treatment arms.

[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.



(a) Men



(b) Women

Fig. 4. Post-treatment student beliefs about own percentile, by treatment status and gender.
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.

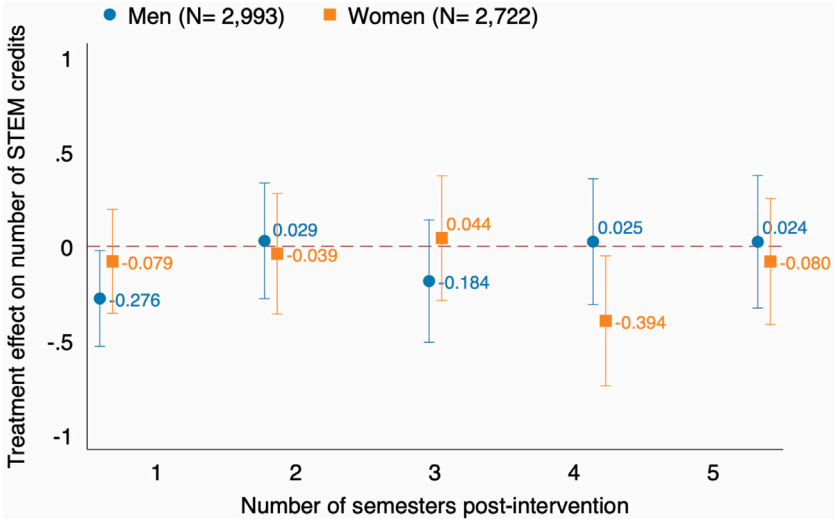
Taken together, the estimated effects of the informational intervention on students' beliefs and subsequent behavior provide limited support for the hypothesis that gender differences in confidence explain different rates of STEM persistence and that information can address the problem. Though the gender differences in beliefs are stark and the intervention does change some students' beliefs, short-term effects of information provision are small, and there are no changes to the longer-term gender gap in major choice. If anything, the point estimates suggest that some women and men may have been discouraged from studying STEM, which has ambiguous welfare implications.

6.5. *Intermediate outcomes and heterogeneity*

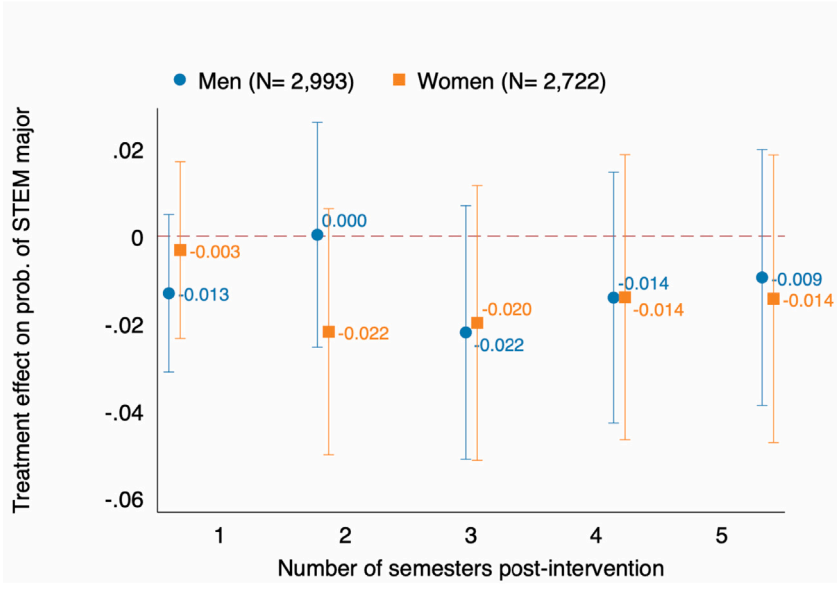
Much of the prior research on feedback provision, in academic and other settings, has focused on effort and performance as an outcome (Ashraf et al., 2014; Azmat et al., 2019; Azmat & Iriberri, 2010; Bandiera et al., 2015; Dobrescu et al., 2019; Goulas & Megalokonomou, 2015; Tran & Zeckhauser, 2012). 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

[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



(a) Effects on Number of STEM Credits



(b) Effects on Probability of Declaring a STEM Major

Fig. 5. Medium- and long-term effects on course-taking and major choice, by gender.
Notes: Treatment effects by gender are 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 (Eq. (2)). Bars show 95% confidence intervals based on robust standard errors. Course-taking and major declaration outcomes are based on University of Michigan administrative data. Number of credits are measured in a given semester (not cumulative). A student is coded as declaring a STEM major if they are declared as a STEM or econ major in the given semester or if they graduated with a degree in a STEM or econ field. A table of estimated effects, standard errors, and control means appears as Appendix Table A.11.

performance for men, women, or students as a whole. Although the point estimates for both final exam and final course performance are negative for men (−0.013 and −0.141, respectively), the lower bounds of the 95 percent confidence intervals imply that men could have at most decreased their final exam and course performance by less than a percentage point, suggesting effort and performance were not a key mechanism through which changing beliefs affected behavior.

The intervention could change students' beliefs about their ability to succeed in STEM, which could serve as an intermediate channel between their beliefs about their performance and their behavior. To measure this, I construct an index capturing students' beliefs about their ability to succeed in STEM, which aggregates responses to items about their grades being "good enough" for STEM, a series of STEM-self-efficacy items, and items about identifying with being a "math person" or "science person".[34] The results are included as the middle left panel of Table 6. The effects of the intervention on this success index are small

allows students to opt out of the final exam (they can drop their lowest exam score, so many choose not to take the final), so I do not include it in my analyses of final exam performance.

[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.

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

By calling attention to grades and academic performance, the intervention may have increased students' academic stress levels, a possible mechanism to explain the somewhat negative effects on course-taking and major choice. 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. (A higher value indicates higher stress.) The middle right panel of Table 6 shows no change to students' stress about grades, overall or by gender.

As additional intermediate outcomes, I examine short-term subjective interest in STEM, measured 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.[35] As shown in the bottom of Table 6, I find small, negative, statistically insignificant effects on subjective STEM intent and small negative effects on STEM interest. The effects on the STEM interest index are negative for both men and women (−0.045 and −0.085 standard deviations, respectively), and the effect is more negative for women. However, both effects are small (less than one tenth of a standard deviation) and I cannot reject that they're equal. This aligns with the negative (though statistically insignificant) effects on ultimate major declaration (see Fig. 5, panel (b)).

Appendix Tables A.13 through A.16 report estimated effects on STEM persistence by a number of pre-treatment characteristics, including student level (first year or sophomore vs. junior or senior), intended major, course subject, instructor gender, 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 behavior, at least in the short term (Appendix Table A.13). 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 and major declaration by the most (Appendix Table A.14). However, by splitting the sample into seven subjects, I don't have the power for subject-by-subject comparisons.

I find no significant differences by instructor gender (Appendix Table A.15) or gender composition of the course (Appendix Table A.16), though this analysis is again underpowered due to the loss of sample size from the cancellation of the second round of the study.

7. Discussion

One of the most striking findings of this study is a descriptive one: men are significantly more overconfident and women more underconfident about their relative performance in STEM courses. 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.

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; Coutts, 2019; Mobius et al., 2014) and that people update less when the information is about a gender-incongruent domain (Coffman et al., 2019), others find the opposite (Goulas & Megalokonomou, 2015; Owen, 2010). The patterns by prior beliefs are broadly but not fully consistent with belief updating, with an overall pattern of initially overconfident students decreasing their STEM persistence in response to bad news. Again, the literature is mixed on asymmetric updating, with some finding people react more strongly to bad news than good news (Coutts, 2019) and others finding the opposite (Mobius et al., 2014).

Though women update in a way suggesting an increase in their relative performance beliefs, they do not become more likely to persist in STEM. If anything, some women (along with some men) may have been discouraged. Understanding why women's choices are largely unmoved is critical to fully understanding gender differences in field choice. 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.

A leading explanation is that women have a comparative advantage in non-STEM, which remains even after revising STEM beliefs (Breda & 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 (Berlin & Dargnies, 2016; Buser et al., 2014; Niederle & Vesterlund, 2011). On the other hand, Fischer (2017) finds that women are more responsive than men to the composition of their peers, with women being less likely to persist in STEM if they are quasi-randomly assigned to an introductory chemistry course with higher-ability peers (and no effect for men). While Fischer's (2017) finding that low-performing students are discouraged by high-ability peers is consistent with the negative effects I find for students receiving bad news, the differences by gender in her study are inconsistent with the lack of gender differences I find.

There are several other explanations for the lack of effects I find, which I cannot fully rule out. First, the information I provided was about relative performance, which by definition is about two things: a student's own performance and that of their peers. Put differently, relative performance feedback also provides a signal about course or major difficulty; learning that a student is doing relatively better than they thought could also be interpreted as learning that a course is

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

more difficult than they thought. Recent evidence suggests that students prefer less difficult majors (Ersoy & Speer, 2023). Good news about an underconfident student's own performance in STEM may be counteracted by bad news about how difficult STEM is. Second, though the intervention was designed to provide information about aptitude in STEM courses specifically, students may have interpreted it as information about general ability. If they revised their beliefs about their general rather than STEM ability, we wouldn't expect major choice to change. Unfortunately, I only measure students' beliefs about their relative STEM performance, so cannot provide evidence for or against this explanation.

Beliefs about oneself and stereotypes about academic subjects and occupations are formed over a person's entire life, with gender differences emerging in children's own beliefs as young as age six (Bian et al., 2017; Cvencek et al., 2011). Students' beliefs, performance, and choices are influenced by their early environments, including the gender stereotypes held by their parents and teachers (Carlana, 2019; Jacobs, 1991). One explanation for my lack of positive effects is that the period of postsecondary education may be too late to correct underconfidence learned over a lifetime, and providing information may backfire by reinforcing stereotypes for lower-performing women. Intervening earlier may be more successful in changing beliefs and behavior.

Features of the intervention itself may explain its lack of more positive effects. Although some of the information students received (e.g., course grades of median STEM majors) was novel, the information as a whole may not add much relative to what they already know via instructors and publicly-available information. Students may have ignored or disregarded the content due to method of delivery – online, through a learning management system, a single time – and may respond better to information delivered in-person, multiple times, and/or coming directly from a trusted source like the course instructor. One month after the intervention, only 39 percent of treated students correctly identified the course median for STEM majors – a statistic they were directly told – suggesting a high degree of inattention.

Given the light touch nature of the intervention and the complex nature of the targeted behavior, a reasonable null hypothesis for effects on major choice may in fact be no change to academic decisions. The choice of college major and subsequent career path is a hugely consequential choice based on preferences and beliefs that have formed over eighteen years prior to entering college. Moreover, many factors matter for major choice, from the large and obvious—e.g., expected earnings and employment (see Patnaik et al. 2021 for a review)—to the seemingly small—e.g., the semester in which a student takes an introductory course (Patterson et al., 2023). And, the various factors likely interact in complex ways. In the current experiment, all of these other factors are held constant, stacking the deck against meaningful behavioral change.

This study sits within a broader body of research that tries to address behavioral barriers in educational decision-making. This literature is somewhat mixed. Some light-touch, informational interventions have proved successful at encouraging behaviors such as FAFSA filing (Page et al., 2020) and college application, enrollment, and persistence (Castleman & Page, 2015, 2016, 2017; Hoxby & Turner, 2013), but some – especially those attempted at scale – have not (Avery et al., 2021; Bergman et al., 2019; Bird et al., 2021; Gurantz et al., 2021). Page et al. (2022) suggest that nudges are most effective when they target acute, time-sensitive tasks (such as filing a form by a deadline) rather than providing more general academic advice or targeting larger decisions. Given these prior findings and the high-stakes nature of major choice, it is perhaps not surprising that the intervention studied here was not more successful. However, I cannot rule out that information about relative performance might be more effective if delivered in person (as in Porter and Serra 2019), if it were more targeted to underrepresented students (as in Bayer et al. 2019), if it were delivered to younger students, or to a different population of college students.

8. Conclusion

Gender differences in college major choice 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 large 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, and that these beliefs are correlated with later behavior. Consistent with theory that beliefs matter for educational choices, providing information may have decreased STEM persistence for students who received bad news. However, students who received good news – in particular, underconfident women – did not display an equivalent increase in persistence, and the overall gender gap in major choice (by five semesters later) was unchanged.

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.

This work speaks to the limits of light-touch interventions in changing consequential behaviors such as major choice. There is a growing consensus in the economics of education literature that "nudge" interventions can be effective at targeting small, self-contained tasks, but that larger behaviors such as college persistence and major choice seem to require more intensive, sustained intervention (Oreopoulos, 2020; Page et al., 2022). A more intensive intervention or one targeting younger students may be effective at changing beliefs and behavior even more, but researchers should design such interventions carefully to avoid discouraging students with bad news or reinforcing stereotypes. Taken in context, my findings suggest that biased beliefs about relative academic performance may be one important piece of the large, complex issue of decisions about major choice and gender differences in STEM. However, increasing women's STEM participation likely requires additional approaches.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data used in the paper come from two restricted-access sources. The first is student-level records from the University of Michigan's Office of Enrollment Management (OEM). These data were obtained through a data sharing agreement between the researcher and OEM. These data are available by request for researchers meeting eligibility criteria, including IRB approval, as outlined at https://enrollment.umich.edu/data/learning-analytics-data-architecture-larc. The second is student-level survey data from the University of Michigan's Center for Academic Innovation (CAI). These data were obtained through a partnership and data sharing agreement between the researcher and CAI. Access to these data require IRB approval and a data use agreement.

Acknowledgments

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. I thank Sue Dynarski, Sara Heller, Kevin Stange, and Charlie Brown for their invaluable support and guidance. Peter Blair, Sarah Cohodes, Ashley Craig, Amanda Griffith, Doug Webber, and Basit Zafar provided helpful comments on early drafts. This work has benefited from numerous conversations with colleagues at the University of Michigan and Colby College. 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, the Liberal Arts College Public and Labor Conference, the Allied Social Science Associations Annual Meeting, and the CESifo Economics of Education Area Conference.

Funding

This work was supported by the U.S. Department of Education's Institute of Education Sciences, PR/Award R305B150012#.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.econedurev.2023.102479.

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., Arcidiacono, P., & Maurel, A. (2016). The analysis of field choice in college and graduate school: Determinants and wage effects. In *Handbook of the economics of education, vol. 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., Aucejo, E., Maurel, A., & Ransom, T. (2016). *College attrition and the dynamics of information revelation*. National Bureau of Economic research working paper 22325.

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

Aucejo, E., & James, J. (2021). The path to college education: The role of math and verbal skills. *Journal of Political Economy*, *129*(10), 2905–2946.

Avery, C., Castleman, B. L., Hurwitz, M., Long, B. T., & Page, L. C. (2021). Digital messaging to improve college enrollment and success. *Economics of Education Review*, *84*, Article 102170.

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

Avitzour, E., Choen, A., Joel, D., & Lavy, V. (2020). *On the origins of gender-biased behavior: The role of explicit and implicit stereotypes*. National Bureau of Economic research working paper 27818.

Aycock, L. M., Hazari, Z., Brewe, E., Clancy, K. B., Hodapp, T., & Goertzen, R. M. (2019). Sexual harassment reported by undergraduate female physicists. *Physical Review Physics Education Research*, *15*(1), Article 010121.

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

Azmat, G., Calsamiglia, C., & Iriberri, N. (2016). Gender differences in response to big stakes. *Journal of the European Economic Association*, *14*(6), 1372–1400.

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

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

Bayer, A., Bhanot, S. P., & Lozano, F. (2019). Does simple information provision lead to more diverse classrooms? Evidence from a field experiment on undergraduate economics. *AEA Papers and Proceedings*, *109*, 110–114.

Bergman, P., Denning, J. T., & Manoli, D. (2019). Is information enough? The effect of information about education tax benefits on student outcomes. *Journal of Policy Analysis and Management*, *38*(3), 706–731.

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

Bettinger, E. P., & Long, B. T. (2005). Do faculty serve as role models? The impact of instructor gender on female students. *AEA Papers and Proceedings*, *95*(2), 152–157.

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

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

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

Bird, K. A., Castleman, B. L., Denning, J. T., Goodman, J., Lamberton, C., & Rosinger, K. O. (2021). Nudging at scale: Experimental evidence from FAFSA completion campaigns. *Journal of Economic Behaviour and Organization*, *183*, 105–128.

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

Booth, A. L., Cardona-Sosa, L., & Nolen, P. (2018). Do single-sex classes affect academic achievement? An experiment in a coeducational university. *Journal of Public Economics*, *168*, 109–126.

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

Bostwick, V. K., & Weinberg, B. A. (2022). Nevertheless she persisted? Gender peer effects in doctoral STEM programs. *Journal of Labor Economics*, *40*(2), 397–436.

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

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

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

Buser, T., Peter, N., & Wolter, S. C. (2017). Gender, competitiveness, and study choices in high school: Evidence from Switzerland. *American Economic Review*, *107*(5), 125–130.

Carlana, M. (2019). Implicit stereotypes: Evidence from teachers' gender bias. *Quarterly Journal of Economics*, *134*(3), 1163–1224.

Carrell, S. E., Page, M. E., & West, J. E. (2010). Sex and science: How professor gender perpetuates the gender gap. *Quarterly Journal of Economics*, *125*(3), 1101–1144.

Castleman, B. L., & Page, L. C. (2015). Summer nudging: Can personalized text messages and peer mentor outreach increase college going among low-income high school graduates? *Journal of Economic Behaviour and Organization*, *115*, 144–160.

Castleman, B. L., & Page, L. C. (2016). Freshman year financial aid nudges: An experiment to increase FAFSA renewal and college persistence. *Journal of Human Resources*, *51*(2), 389–415.

Castleman, B. L., & Page, L. C. (2017). Parental influences on postsecondary decision making: Evidence from a text messaging experiment. *Educational Evaluation and Policy Analysis*, *39*(2), 361–377.

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

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

Coffman, K. B., Collis, M., & Kulkarni, L. (2019). *Stereotypes and belief updating*. Harvard Business School, Working paper.

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

Cvencek, D., Meltzoff, A. N., & Greenwald, A. G. (2011). Math–gender stereotypes in elementary school children. *Child Development*, *82*(3), 766–779.

Delaney, J. M., & Devereux, P. J. (2021). The economics of gender and educational achievement: Stylized facts and causal evidence. In *Oxford research encyclopedia of economics and finance*.

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

Ersoy, F., & Speer, J. D. (2023). *Opening the black box of college major choice: Evidence from an information intervention*. Working paper.
Exley, C. L., & Kessler, J. B. (2022). The gender gap in self-promotion. *Quarterly Journal of Economics*, *137*(3), 1345–1381.
Fischer, S. (2017). The downside of good peers: How classroom composition differentially affects men's and women's STEM persistence. *Labour Economics*, *46*, 211–226.
Franco, C. (2019). *How does relative performance feedback affect beliefs and academic decisions?* Working paper.
Gelbach, J. B. (2016). When do covariates matter? And which ones, and how much? *Journal of Labor Economics*, *34*(2), 509–543.
Gonzalez, N. (2017). *How learning about one's ability affects educational investments: Evidence from the advanced placement program*. Mathematica policy research working paper 52.
Goulas, S., & Megalokonomou, R. (2015). *Knowing who you are: The effect of feedback information on short and long term outcomes*. Working paper.
Griselda, S. (2022). *The gender gap in math: What are we measuring?* Working paper.
Gurantz, O., Howell, J., Hurwitz, M., Larson, C., Pender, M., & White, B. (2021). A national-level informational experiment to promote enrollment in selective colleges. *Journal of Policy Analysis and Management*, *40*(2), 453–479.
Hakimov, R., Schmacker, R., & Terrier, C. (2022). *Confidence and college applications: Evidence from a randomized intervention*. Technical report, WZB discussion paper.
Hoxby, C., & Turner, S. (2013). Expanding college opportunities for high-achieving, low income students. *Stanford Institute for Economic Policy Research Discussion Paper*, *12*(014), 7.
Hsieh, C.-T., Hurst, E., Jones, C. I., & Klenow, P. J. (2019). The allocation of talent and US economic growth. *Econometrica*, *87*(5), 1439–1474.
Iriberri, N., & Rey-Biel, P. (2021). Brave boys and play-it-safe girls: Gender differences in willingness to guess in a large scale natural field experiment. *European Economic Review*, *131*, Article 103603.
Jacobs, J. E. (1991). Influence of gender stereotypes on parent and child mathematics attitudes. *Journal of Educational Psychology*, *83*(4), 518.
Jensen, E. J., & Owen, A. L. (2000). Why are women such reluctant economists? Evidence from liberal arts colleges. *American Economic Review*, *90*(2), 466–470.
Kling, J. R., Liebman, J. B., & Katz, L. F. (2007). Experimental analysis of neighborhood effects. *Econometrica*, *75*(1), 83–119.
Kugler, A. D., Tinsley, C. H., & Ukhaneva, O. (2021). Choice of majors: Are women really different from men? *Economics of Education Review*, *81*, 1–19.
Kuhn, A., & Wolter, S. C. (2022). Things versus people: Gender differences in vocational interests and in occupational preferences. *Journal of Economic Behaviour and Organization*, *203*, 210–234.
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., Fox, P. W., & Punćcohaŕ, J. (1994). Highly confident but wrong: Gender differences and similarities in confidence judgments. *Journal of Educational Psychology*, *86*(1), 114.
Marshman, E. M., Kalender, Z. Y., Nokes-Malach, T., Schunn, C., & Singh, C. (2018). Female students with A's have similar physics self-efficacy as male students with C's in introductory courses: A cause for alarm? *Physical Review Physics Education Research*, *14*(2), Article 020123.
Minnotte, K. L., & Pedersen, D. E. (2023). Sexual harassment, sexual harassment climate, and the well-being of STEM faculty members. *Innovative Higher Education*, 1–18.
Mobius, M. M., Niederle, M., Niehaus, P., & Rosenblat, T. S. (2014). *Managing self-confidence*. Working paper.
Niederle, M., & Vesterlund, L. (2007). Do women shy away from competition? Do men compete too much? *Quarterly Journal of Economics*, *122*(3), 1067–1101.
Niederle, M., & Vesterlund, L. (2011). Gender and competition. *Annual Review of Economics*, *3*(1), 601–630.
Olson, S., & Riordan, D. G. (2012). *Engage to Excel: Producing one million additional college graduates with degrees in science, technology, engineering, and mathematics*. ERIC, Report to the President, Executive Office of the President.
Oreopoulos, P. (2020). *Promises and limitations of nudging in education*. IZA discussion paper no. 13718.
Owen, A. L. (2010). Grades, gender, and encouragement: A regression discontinuity analysis. *The Journal of Economic Education*, *41*(3), 217–234.
Owen, A. L., & Hagstrom, P. (2021). Broadening perceptions of economics in a new introductory economics sequence. *The Journal of Economic Education*, *52*(3), 175–191.
Page, L. C., Castleman, B. L., & Meyer, K. (2020). Customized nudging to improve FAFSA completion and income verification. *Educational Evaluation and Policy Analysis*, *42*(1), 3–21.
Page, L. C., Meyer, K., Lee, J., & Gehlbach, H. (2022). *Conditions under which college students can be responsive to nudging*. Annenberg Institute at Brown University EdWorkingPaper no. 20-242.
Page, L., & Ruebeck, H. (2022). Childhood confidence, schooling, and the labor market: Evidence from the PSID. *Journal of Human Resources* (in press; published online September 2022).
Patnaik, A., Wiswall, M., & Zafar, B. (2021). College majors. In *The Routledge handbook of the economics of education* (pp. 415–457). Routledge.
Patterson, R. W., Pope, N. G., & Feudo, A. (2023). Timing matters: Evidence from college major decisions. *Journal of Human Resources*, *58*(4), 1347–1384.
Porter, C., & Serra, D. (2019). Gender differences in the choice of major: The importance of female role models. *American Economic Journal: Applied Economics*, *12*(3), 226–254.
Roy, A. D. (1951). Some thoughts on the distribution of earnings. *Oxford Economic Papers*, *3*(2), 135–146.
Speer, J. D. (2023). Bye Bye Ms. American Sci: Women and the leaky STEM pipeline. *Economics of Education Review*, *93*, Article 102371.
Tran, A., & Zeckhauser, R. (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 (2022). Learning analytics data architecture (LARC) data set.
Vincent-Ruz, P., Binning, K., Schunn, C. D., & Grabowski, J. (2018). The effect of math SAT on women's chemistry competency beliefs. *Chemistry Education Research and Practice*, *19*(1), 342–351.
Webber, D. A. (2019). *Projected lifetime earnings by major*. Technical report.
Wiswall, M., & Zafar, B. (2015). Determinants of college major choice: Identification using an information experiment. *Review of Economic Studies*, *82*(2), 791–824.
Wozniak, D., Harbaugh, W. T., & Mayr, U. (2014). The menstrual cycle and performance feedback alter gender differences in competitive choices. *Journal of Labor Economics*, *32*(1), 161–198.
Xue, Y., & Larson, R. C. (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.