Essays on Gender and Education

by

Maria Paola Ugalde Araya

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved April 2023 by the
Graduate Supervisory Committee:

Esteban Aucejo, Co-Chair
Basit Zafar, Co-Chair
Tomas Larroucau

ARIZONA STATE UNIVERSITY

May 2023

ABSTRACT

This dissertation studies the differences in how men and women react to feedback or information about their performance in educational settings and how these differences might impact women's decisions to stay away from traditionally male-dominated fields.

The first chapter analyzes the gender differences in reaction to low performance during high school. I focus on the decision of North Carolina public high school students to enroll in advanced math or English classes after learning about their performance on statewide standardized tests in each subject. I find that women are more responsive to low-performing than men. Women that perform poorly on their tests are less likely than their higher-performance peers to enroll in advanced classes, while men's likelihood is the same regardless of performance.

It has been documented that the probability of women continuing their studies in male-dominated fields – like Science, Technology, Engineering, and Mathematics (STEM) and business – is more sensitive to their performance in relevant courses at the beginning of college relative to men. The second chapter studies these gender differences in grade sensitivity during college. Using novel survey data, I estimate students' sensitivity to grades and find that women value an extra grade point average (GPA) unit more than men. I find that anticipated discrimination in the labor market of male-dominated fields is important to understand this gender gap in grade sensitivity. I further provide evidence of the gender differences in beliefs about labor market discrimination in different fields.

The last chapter investigates the dynamic effects of feedback in an experimental setting. I explore how individuals update their beliefs and choices in response to good or bad news over time in two domains: verbal skills and math. I find significant gender gaps in beliefs and choices before feedback: men are more optimistic about their performance and more willing to compete than women in both domains, but the gaps are significantly larger in math. Feedback significantly shifts individuals' beliefs and choices immediately after receiving it. However, there is substantial persistence of gender gaps over time. This is particularly true among the set of individuals who receive negative feedback.

DEDICATION

A papi, a mami y a Dani, por creer en mí.

A abuelo Juan, el abuelo más orgulloso de sus nietas.

ACKNOWLEDGEMENTS

I am especially grateful to Basit Zafar, and Esteban Aucejo for their guidance, support, and encouragement. I would also like to thank Tomas Larroucau and Dan Silverman for their time and advice. I am profoundly grateful to Zachary Tobin, Jacob French, Kun Zhang, Jakob Dowling, Juan Odriozola, and Minju Jeong for stimulating discussions and careful insights.

I am particularly thankful to Zachary Tobin, Laura Chavez, and Diego Aguiar for providing the support, confidence, and outside perspective that I needed through the countless crises that the PhD journey represented. I could not have done it without them.

Finally, I want to thank my parents, Nelly and Juan, and my sister Dani, for their love, understanding, and unwavering confidence in me. I am grateful to my grandparents, aunts, uncles, and cousins that kept me in their prayers and thoughts. Special thanks to Abuela Cecilia for always lighting a candle when I need it.

TABLE OF CONTENTS

Page

APPENDIX

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

Figure Page

Chapter 1

LOW PERFORMANCE IN MATH AND ENGLISH: DO WOMEN REACT DIFFERENTLY THAN MEN?

1.1 Introduction

Women are underrepresented in high-level positions in many areas. They represented 31% of the senior management positions in 2021 (Grant Thornton International Ltd, 2021) and held only 7.4% of the CEO positions in the Fortune 500 companies in 2020 (Pew Research Center, 2021). Making it to the top requires people to overcome setbacks at every stage in life, from low grades and college admission rejections, to bad job interviews or performance evaluations. If women react more adversely to these setbacks than men, an early defeat might preclude a woman from advancing.

In this chapter, I study the gender difference in reaction to low performance. I focus on the decision of high school students to enroll in advanced math or English classes after learning about their performance on subject-specific standardized tests. The idea is to determine whether women react significantly more to low performance on those tests than men, and if the reaction varies with the subject. Specifically, I establish if low performance differentially impacts the probability that a woman or a man enrolls in advanced classes.

There is a growing literature that concludes that women are more likely to drop out of male-dominated majors like Economics and STEM (Science, Technology, Engineering, Mathematics) after experiencing a difficulty early in their college careers. In Goldin (2015), the author documents that women getting B- or less in their introductory economics course are less likely to graduate with an economics major than men with the same final grades. Similarly, Rask and Tiefenthaler (2008) estimate that the probability of a woman enrolling in a second course in economics decreases if she gets grades in the lowest quartile of the distribution during her first economics class, but the same is not true for men. In Kugler *et al.* (2021), the authors conclude that if women believe that men are inherently a better fit for STEM majors, and if there are less women than men in their major, they are more likely to perceive their low grades as confirmation of their unfitness for their male-dominated STEM major and drop out. Further, Ahn *et al.* (2019) estimate a structural model in which women care more about grades and conclude that if women had the same preferences for grades as men the gender gap in STEM majors would close by 8%. However, a question that remains open is whether the reaction to setbacks is the same in male and female-dominated fields. Furthermore, given that most of the existing evidence is at the higher education level, analyzing the reaction to setbacks earlier in life can provide insight about critical decisions that occur later on.

The interest in the difference between subjects lies in the common association of men with math and women with English and languages. For example, according to the National Center for Education Statistics (2018), in 2015 more male than female students identified math as their favorite subject, and more female than male students reported reading as their favorite activity. Riegle-Crumb and Humphries (2012) concludes that teachers believe math is easier for white boys than for white girls. Women are less confident in their math abilities (Ellis *et al.*, 2016, Ganley and Lubienski, 2016) and have a larger advantage over men in verbal skills (Aucejo and James, 2021; Delaney and Devereux, 2019; Breda and Napp, 2019). However, taking more high school math courses increases wages for female college graduates and the likelihood of a woman majoring in “more technical and nontraditional fields” (Levine and Zimmerman, 1995)[1].

I study the course-taking behavior of North Carolina public high school students for the cohorts starting 9^{th} grade in the 2013 and 2014 scholar years. The North Carolina Department of Public Instruction (NCDPI) uses standardized End-of-Course (EOC) tests to assess students' knowledge about specific subjects for high school accountability purposes (North Carolina Department of Public Instruction, 2019b). During high school, every student must take an EOC test at the end of the first math course (Math 1), the second English course (English 2) and the Biology course. For each subject tested, there is a threshold score above which students are considered proficient, i.e. they have sufficient command of skills for the respective course and are prepared for further studies in the subject (North Carolina Department of Public Instruction, 2016a). I focus on the EOC tests for Math 1 and English 2, and the corresponding choice of regular versus advanced class in Math 2 and English 3.

In North Carolina, each spring semester students make their choices for the classes they will be taking the next year. Often, students decide between the regular or honors version of a class, and when available they can consider taking Advanced Placement (AP) and International Baccalaureate (IB) classes. These advanced classes are more difficult (AP and IB are college-level classes), and cover more topics than the regular class. The benefits associated with taking advanced classes during high school are numerous. For example, taking advanced courses is related to increases in the probability of high school graduation, 4-year college attendance and college graduation, and with the completion of more college credits and a higher GPA during college (Long *et al.*, 2012). Moreover, taking advanced math classes has a positive effect on earnings later in life (Rose and Betts, 2004; Joensen and Nielsen, 2009).

[1] In Levine and Zimmerman (1995), a nontraditional female field is a major in which more than 70% of the students are men.

Therefore, it is important to understand if the advanced course-taking behavior of high school students of different genders is influenced differently by the perception of low performance on a test.

The existence of the proficiency cutoffs on the North Carolina EOC tests suggests the implementation of a regression discontinuity (RD) design. However, the main interest is not in the discontinuity generated by the cutoff that quantifies the effect of low performance on the likelihood of enrolling in an advanced class, but in the gender difference in the discontinuity. Thus, I use a difference-in-discontinuity approach that combines regression discontinuity (RD) and differences-in-differences (DD) in order to identify the gender differences in the discontinuity created by the cutoffs.

I find that the effect of low performance is not the same for math and English. Women react more to low performance in math than in English relative to men. The probability of enrolling in honors Math 2 changes with performance for women, but not for men. Women that perform poorly on the Math 1 EOC tests are 5 to 7 percentage points less likely to enroll in honors Math 2 while men's likelihood is the same regardless of performance status. On the other hand, the effect of low-performance on the English test is smaller than in math. As in the math case, men do not seem to react to low-performance. However, women's probability of enrolling in Advanced English 3 is between 2 and 4 percentage points lower among non-proficient students. The fact that women react more to low performance in math than in English suggests that the area in which a woman is facing a difficulty might be relevant for her reaction. This result is in line with Kaganovich *et al.* (2021), which concludes that women's grade sensitivity depends on the area or category.

My primary results focus on advanced course-taking decisions for Math 2 and English 3, however I also analyze the effect of low performance on higher-level advanced course-taking behavior at the end of high school. I find that the number of higher-level advanced courses taken by non-proficient women is smaller relative to non-proficient men. Roughly, these differences translate to a reduction of 6% and 14% in the number of higher-level advanced courses taken by women in English and math, respectively. These results are in line with Tan (2020) which finds that college students at the margin of two letter grades that receive the worse grade are more likely to take easier classes in the subsequent years.

There is evidence that test scores can influence students' decisions about post-secondary education (Papay *et al.*, 2016). Therefore, I study the effect of low performance on the plans to attend a 4-year college.[2] I find that women are 2 and 3 percentage points less likely to make plans of going to college than non-proficient men when deemed not proficient on the English 2 and Math 1 EOC tests, respectively.

My findings are consistent with women interpreting their non-proficiency status as a discouraging signal of their ability, and this may affect their future behavior in terms of class enrollment and college attendance. There is evidence that women attribute negative feedback to lack of ability, update their beliefs more pessimistically, and are less willing to compete than men (Roberts and Nolem-Hoeksema, 1989; Shastry *et al.*, 2020; Berlin and Dargnies, 2016; Buser and Yuan, 2019a). All of these traits could lead to women avoiding "harder" classes after performing poorly on the standardized tests. Additionally, the fact that the effect of low performance is stronger in math than in English highlights the importance of the stereotype associated with a given domain, which is in line with results from laboratory settings in which women and men respond to feedback differently depending on the gender stereotype associated with a task (Kiefer and Shih, 2006; Coffman *et al.*, 2019).

[2] I only observe what students plan to do after high school, but I do not know if they executed those plans.

1.2 Institutional Background and Data

The North Carolina Department of Public Instruction (NCDPI) uses standardized end-of-course (EOC) tests to "sample a student's knowledge of subject-related concepts as specified in the North Carolina Standard Course of Study and to provide a global estimate of the student's mastery of the material in a particular content area" (North Carolina Department of Public Instruction, 2019b) for high school accountability purposes. During high school, every student must take an EOC test at the end of the first math course (Math 1), the second English course (English 2) and the Biology course. I focus on Math 1 and English 2 tests, usually taken during 9^{th} and 10^{th} grade, respectively. Each of these tests represent 20% of the final course grade (North Carolina Department of Public Instruction, 2016b, 2016c).

Based on the test scores, students are classified into five achievement levels, where one is the lowest achievement level and five the highest. Students that receive at least an achievement level of 3 are considered proficient, which means that they have sufficient command of skills for the respective course and are prepared for further studies in that subject (North Carolina Department of Public Instruction, 2016a). For each subject, there is an established threshold above which students are considered proficient. These clear cutoffs allow for the application of the empirical strategy described in Section 1.3.

Most high schools in North Carolina employ a block schedule or semester plan, in which students take four classes each semester, for a total of eight per year (Averett, 1994). High school graduation requirements in North Carolina include the completion of four math credits and four English credits, which are equivalent to four courses in each subject, usually taken one each year of high school.[3] The block schedule and

[3] Students can take more than the four required courses in each subject.

adequate progress from grade to grade make it possible to study each subject only one semester each high school year, for instance, by taking English during fall semesters and math during spring semesters.[4]

Each scholar year, during the spring semester (around March or April), students make the choices for the classes they will be taking the next year. When choosing the classes, students often must decide whether to enroll in a regular or honors version of a given course. This is the case for Math 2, for instance. The honors version is a class with a higher level of difficulty, which studies the topics in a deeper way than the regular version and sometimes covers more topics. Given the higher level of difficulty, students taking honors classes get quality points that make grades from honors classes have a higher weight in GPA calculations. For example, a C in a honors class is equivalent to a B in a regular class (North Carolina Department of Public Instruction, 2022).

Once students advance to higher grades, the available options can be more than just honors and regular. For example, the NCDPI allows for Advanced Placement (AP) classes in English Language and Literature or, if offered, an International Baccalaureate (IB) English class to count as the third and fourth English credits. AP and IB courses are college-level classes in which students can earn college credit depending on performance on a test at the end of the course. High school students get even more quality points for taking AP and IB courses than for honor classes.

However, quality points and more knowledge are not the only benefits of taking advanced classes. For instance, in Long *et al.* (2012), the authors conclude that taking

[4] However, this is not the only possibility, the block schedule and the option to take high school level classes during middle school allow for different paths. For example, four math credits can be obtained by taking (and passing) one math class each semester during the first two years of high school.

more rigorous courses, in math, English or science, during high school increases the probability of high school graduation and 4-year college attendance.[5] Additionally, college students who took advanced classes during high school tend to complete more college credits, have a higher GPA, and a higher probability of graduation. Moreover, studies like Rose and Betts (2004) and Joensen and Nielsen (2009) find that taking advanced math classes during high school has a positive effect on earnings later in life.[6]

In order to determine if there is a gender difference in the reaction to low performance on EOC tests, I use administrative records from the North Carolina Education Research Data Center (NCERDC). My analysis focuses on the cohorts that began public high school on the fall of 2013 and 2014. Since one of my main objectives is to study the effect of test scores on class choices for the next scholar year, I restrict the samples to those students taking the relevant EOC tests during the fall semesters. In this way I guarantee that students have received the exam results prior to making course decisions. Additionally, I restrict the samples to students for which the relevant transcript information for the following school year is observed.

The Math 1 sample includes 27,997 students that began high school in the 2013-2014 or 2014-2015 school years and took the Math 1 EOC test at the end of the fall semester during their freshman year.[7] The English 2 sample includes 72,395

[5] The National Center for Education Statistics (2019) report about math, science and reading instruction finds a positive correlation between taking advanced math classes and 4-year college acceptances.

[6] In Joensen and Nielsen (2009), the increase in earnings is due to a higher probability of attending college.

[7] About 16% of the students took the Math 1 class during the fall of their freshman year, 51% took it after the fall semester and 33% took it during 8th grade (some middle schools offer 9^{th} grade-level math classes, which allows students to gain high school credits and take 10^{th} classes

Table 1.1: Sample Characteristics

	Math Sample	English Sample
Female	0.51	0.51
Black	0.26	0.24
Hispanic	0.13	0.11
White	0.54	0.58
EDS	0.51	0.42
Non-Prof. Middle School Math test	0.63	0.35
Non-Prof. Middle School Reading test	0.53	0.38
Non-Prof. Math 1 test	0.49	-
Non-Prof. English 2 test	-	0.35
N	27,997	72,395

Note: Table presents sample proportions of variables of interest. Math sample includes students that took the Math 1 class during the fall semester of their freshman year. English sample includes students that took the English 2 class during the fall semester of their sophomore year. Non-Prof.: non-proficiency, EDS: Economically Disadvantaged Student. *Significant at 10%, **5%, ***1%.

students that began 10th grade in the fall of 2013 or 2014 and that took the English 2 EOC test at the end of that semester.[8] Table 1.1 shows summary statistics for both samples. They are balanced in terms of gender with 51% of women. More than half the students are white (54% math, 57% English), around 25% Hispanic and 12% or less are black. The proportion of economically disadvantaged students (EDS) is 52% and 43% for the math and English samples, respectively. In terms of academic outcomes, 63% (35%) of the students in the math (English) sample were deemed to proficient on their middle school math test. On the middle school reading test these numbers are 53% and 38% for the math and English samples, respectively.[9] Finally,

during their freshman year). Fall students performed better in academic terms than the after fall, but worse than the students that took the class during middle school. See Table A.1.

[8] Around 38% of the students took the English 2 class during the fall of their sophomore year, 62% took it after the fall semester and 7% during their freshman year. There are not economically significant differences between the fall and after fall students. See Table A.2

[9] The reading and math tests during 8^{th} are part of the end-of-grade (EOG) exams that North Carolina students take at the end of the year from grades 3 to 8 in order to measure their performance

there is a sizable proportion of students that performed poorly on the tests of interest in the respective samples: 49% of the math sample was deemed not proficient on the Math 1 test, and 35% of the students in the English sample were not proficient on the English 2 test.

Additionally, I am interested in the effect of low performance in outcomes closer to the end of high school like the number of higher-level advanced classes in math and English that students take, and plans to attend a 4-year higher education institution. Therefore, I further restrict the samples to students that graduated at the end of their fourth year for this part of the analysis.[10]

1.3 Empirical Strategy

Given that the main objective is to determine if there is a gender difference in the reaction to low performance on EOC tests, and the existence of a cutoff below which students are considered non-proficient in that subject, I use a difference-in-discontinuity approach that combines regression discontinuity (RD) and differences-in-differences (DD).[11] The RD part exploits the cutoff and requires the identifying assumption that students just above and just below the cutoff are very similar except for the difference in proficiency status. The RD design estimates the size of a discontinuity, in other words it identifies the effect of low performance on an EOC test on the likelihood of enrolling in the advanced version of the class for the next school year. The DD part allows me to determine if there is a gender difference in

on "the goals, objectives, and grade-level competencies specified in the North Carolina Standard Course of Study" (North Carolina Department of Public Instruction, 2019a).

[10] The main results are robust to use the restricted samples.

[11] See Buser and Yuan (2019a), Grembi *et al.* (2016); Eggers *et al.* (2018); Galindo-Silva *et al.* (2020) for examples of this design.

the discontinuity, i.e. if there is a gender difference in the effect of low performance.

This difference in the discontinuity can be estimated by the interaction of the treatment (being non-proficient) and a gender variable. In practice, I estimate the following model:

$$\begin{aligned} Y_{ijkt} = & \beta_0 + \beta_1 \mathrm{F}_i + \beta_2 \text{Non-Prof}_{ijkt} + \beta_3 (\mathrm{F}_i \cdot \text{Non-Prof}_{ijkt}) \\ & + f(S_{ijkt}) + \text{Non-Prof}_{ijkt} \cdot f(S_{ijkt}) + \gamma \mathbf{X}_{ij} + \eta_j + \nu_t + \epsilon_{ijkt} \end{aligned} \tag{1.1}$$

where $k \in \{\text{math, English}\}$, and Y_{ijkt} is the outcome variable for student i at high school j that took the test for subject k during school year t. F_i is an indicator variable equal to one for females. Non-Prof_{ijkt} is an indicator that takes value one when the student is not proficient on the EOC test. η_j and ν_t are high school and year fixed effects, respectively. $\mathbf{X}$ includes controls like middle school test scores, race, EDS. $f(S_{ijkt})$ represents a function of the EOC test score, which is the running variable in this case. The interaction between $f(S_{ijkt})$ and Non-Prof_{ijkt} allows for different slopes above and below the cutoff. The main results in the next sections are estimated using $f(S_{ijkt})$ as a first-degree polynomial, however all the results are robust to using a second-degree polynomial instead. The parameter of interest is β_3, the coefficient on the interaction between being a female and non-proficient that estimates the gender difference in the discontinuity. In all regressions, I follow the recommendation of Lee and Card (2008) for regression discontinuity with a discrete assignment variable and cluster the standard errors at the test score level.

1.3.1 Validity of Design

Given that RD is part of the difference-in-discontinuity design, it is important to make sure that the running variable is exogenous, in other words that test scores are not precisely manipulated. Additionally, all other factors that play a role when deciding whether to take the advanced or regular version of a class are continuously related with the test scores, but more importantly they do not change differentially across genders. These assumptions guarantee that groups of students above and below the proficiency threshold are similar in terms of relevant outcomes and characteristics.

Test Score Manipulation

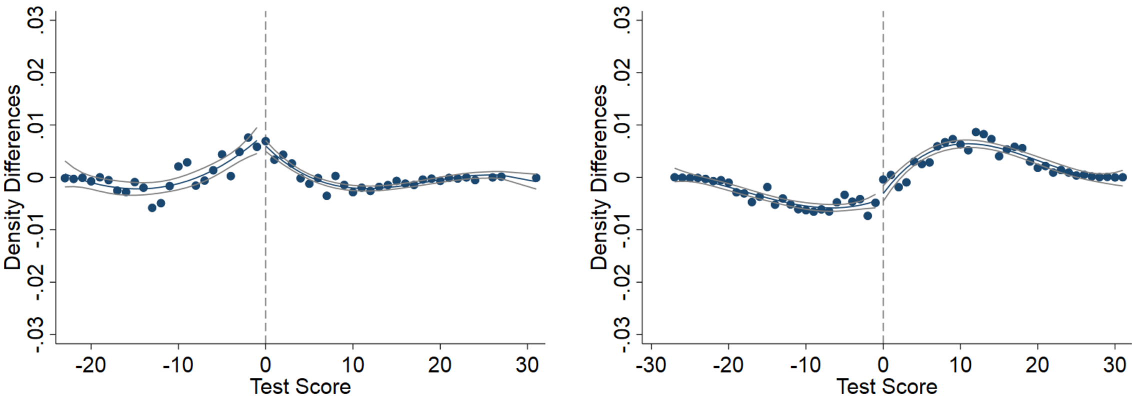
An exogenous running variable is required because manipulation of the test scores can lead to identification problems. I apply the tests described in McCrary (2008) and Frandsen (2017) where the idea is to test the continuity of the running variable density at the cutoff. For both subjects and both genders, the hypothesis of manipulation cannot be rejected at usual significance levels, regardless of the test.[12] However, it is important to remember that the validity of the design is only compromised when the agents can "precisely" manipulate the running variable (Roberts and Whited, 2013), and this kind of manipulation is dubious in the case of the EOC tests in North Carolina.

For instance, one can think that students have some control over their scores but they cannot predict them with certainty or marginally change them around the cutoff. There are several versions of the test, and the exact number of correct answers required

[12] These results are not surprising given the distribution of the test scores in Figure A.1. In both cases, there are spikes in the distribution around the cutoff. However the spikes seem to be similar across genders in both subjects, which suggests that if any manipulation exists, it is similar for men and women. Another explanation for the shape of the distributions is the performance measurement system that was in place in North Carolina at the time. Under that system, high school performance depended heavily on the percentage of students considered proficient on the EOC tests (North Carolina Department of Public Instruction, 2015). Such proficiency-count systems create strong incentives for the teachers to direct resources and attention to students on the margin of being above the proficiency threshold, and to pay less attention to the tails of the distribution (Macartney *et al.*, 2018).

to achieve proficiency is not public information and varies by test version. Given that the test scores are used to assess teacher and high school performance, one might be concerned that the teachers are manipulating the results. However, this seems very implausible because tests are not graded by the professor. Instead the NCDPI has in place a centralized grading system with a rigorous protocol to ensure the security of the materials before and during the tests and to avoid any manipulation when transporting and scanning the answer sheets at the NCDPI offices (North Carolina Department of Public Instruction, 2016b, 2016c).[13]

Figure 1.1: Test Score Densities Differences



(a) Math 1.

(b) English 2.

Note: This is a visual test of the continuity of the difference between female and male densities at the cutoff. The central lines are third-degree local polynomials fitted to the data separately above and below the threshold, the lateral lines represent 95% confidence intervals. Scores are normalized such that a score of 0 or more means proficiency.

On the other hand, given that the interest is in the gender difference in the discontinuity, it is important to guarantee that if any manipulation exists, it is similar across genders. Following the visual test proposed by Grembi *et al.* (2016), Figure 1.1 plots the difference between male and female test score densities for Math 1 and

[13] Dee *et al.* (2019) documents the elimination of teacher test score manipulation once a centralized grading system was adopted in New York.

English 2, along with a third-degree polynomial fitted to the data separately above and below the threshold and 95% confidence intervals. These figures support the assumption that the difference in the densities across genders is continuous at the cutoff.

Continuity of Predetermined Covariates

Given that the main interest is the gender difference in the discontinuity, it is important to rule out the possibility of discontinuities in the predetermined covariates varying by gender. In order to do so I estimate model (1.2) for each subject $k \in$ {math, English} with each of the predetermined controls included in $\mathbf{X}$ as dependent variables, within the optimal bandwidth for each subject used in the main results (±6 for math and ±4 for English).

$$\begin{aligned} X_{ijk} =& \beta_0 + \beta_1 \mathrm{F}_i + \beta_2 \text{Non-Prof}_{ijk} + \beta_3 (\mathrm{F}_i \cdot \text{Non-Prof}_{ijk}) \\ & + f(S_{ijk}) + \text{Non-Prof}_{ijk} \cdot f(S_{ijk}) + \epsilon_{ijk} \end{aligned} \tag{1.2}$$

Results are presented in Tables 1.2 and 1.3 for math and English, respectively. There are no gender differences for any of the math covariates, and the gender difference in the proportion of black students in the English 2 sample is small (2%), however the main results control for all these covariates in order to avoid any biases due to the discontinuities.

The continuity of the covariates, ignoring the possibility of gender differences, is studied by estimating instead model (1.3) for each subject $k \in$ {math, English} with each of the predetermined controls included in $\mathbf{X}$ as dependent variables.

$$X_{ijk} = \beta_0 + \beta_1 \text{Non-Prof}_{ijk} + f(S_{ijk}) + \text{Non-Prof}_{ijk} \cdot f(S_{ijk}) + \epsilon_{ijk} \tag{1.3}$$

Table 1.2: Continuity of the Covariates for Math 1 Test by Gender

	Middle School Math test	Middle School Reading test	Black	Hispanic	EDS
	(1)	(2)	(3)	(4)	(5)
Female (F)	-0.030***	0.150***	0.015	0.009	-0.001
	(0.006)	(0.018)	(0.013)	(0.007)	(0.009)
Non-Prof.	-0.008	-0.010	-0.006	0.018	-0.005
	(0.024)	(0.021)	(0.010)	(0.013)	(0.011)
F*Non-Prof.	-0.017	0.019	0.012	-0.020	0.012
	(0.013)	(0.027)	(0.013)	(0.012)	(0.015)
Mean	-0.10	-0.07	0.24	0.13	0.51
N	14,184	14,152	14,918	14,918	14,918

Note: Dependent variable indicated at the top of each column. Each column follows the same specification: dependent variable regressed on a variable equal to one when the student is a woman, a variable equal to one when the student is deemed non-proficient on the Math 1 EOC test, the interaction of those two and a first degree polynomial of the test score with flexible slopes above and below the proficiency cutoff and high school and year FE. Test scores are standardized such that mean is zero and standard deviation is one. EDS: Economically Disadvantaged Student. Bandwidth of ±6 around the cutoff. Standard errors are clustered at test score level. *Significant at 10%, **5%, ***1%.

Table 1.3: Continuity of the Covariates for English 2 Test by Gender

	Middle School Math test	Middle School Reading test	Black	Hispanic	EDS
	(1)	(2)	(3)	(4)	(5)
Female (F)	-0.184***	-0.101***	0.047***	0.002	0.045**
	(0.016)	(0.009)	(0.008)	(0.007)	(0.014)
Non-Prof.	-0.025	-0.000	0.001	-0.011	-0.016
	(0.040)	(0.013)	(0.006)	(0.011)	(0.009)
F*Non-Prof.	-0.013	0.002	-0.022**	0.007	0.015
	(0.020)	(0.011)	(0.009)	(0.008)	(0.014)
Mean	-0.15	-0.19	0.28	0.13	0.49
N	19,963	19,966	21,472	21,472	21,472

Note: Dependent variable indicated at the top of each column. Each column follows the same specification: dependent variable regressed on a variable equal one when the student is a woman, a variable equal to one when the student is deemed non-proficient on the English 2 EOC test, the interaction of those two and a first degree polynomial of the test score with flexible slopes above and below the proficiency cutoff, and high school and year FE. Test scores are standardized such that mean is zero and standard deviation is one. EDS: Economically Disadvantaged Student. Bandwidth of ±4 around the cutoff. Standard errors are clustered at test score level. *Significant at 10%, **5%, ***1%.

The results are presented in Figures A.2 and A.3 in the Appendix for math and English, respectively. They suggest that there are no statistically significant discontinuities at the cutoff except for the proportion of black students in the English sample (in line with the results of the previous analysis).

Overall, I find no evidence of economically significant discontinuities in covariates at the threshold for both math and English tests, which provides some evidence that supports the chosen empirical strategy and the results presented in the next section.

1.4 Early High School Outcomes

This section presents the main results of the study. Discussion centers on the estimation of model (1.1) where the outcomes of interest are the decisions to enroll in advanced math or English classes the following year after taking the math or English EOC test, respectively. First, I analyze how the gender difference changes when adding different controls for a specific bandwidth and identify the preferred specification. Then, I examine how the main effect changes (or not) when varying the bandwidths in the preferred specification.

1.4.1 Math

Table 1.4 shows the results of the estimation of model (1.1) for math within a bandwidth of ±6 points above and below the proficiency cutoff. The optimal bandwidth is between 6 and 7, according to the procedures suggested by Imbens and Kalyanaraman (2012) and Calonico *et al.* (2019).[14] The outcome variable equals one when the student enrolls in honors Math 2, zero otherwise. The coefficient of interest is the interaction between being a woman and getting a non-proficient score on the Math 1 EOC test (F*Non-Prof). This coefficient estimates the effect of being a non-

[14] The results for a bandwidth of ±7 can be found in Appendix Table A.4.

proficient woman on the likelihood of enrolling in honors Math 2. In other words, it estimates the gender difference in the discontinuity.

The differences across specifications in Table 1.4 are the controls included in each case. The sample size differences arise because it was not possible to recover the middle school test scores for all the students in the math sample. Nonetheless, the proportion of students enrolling in honors Math 2 is around 23% across all specifications.

First, note that the coefficient for being non-proficient is not statistically different from zero and small for most of the specifications. This suggests that the probably of men enrolling in honors Math 2 does not change due to the proficiency level they receive on the Math 1 EOC test. On the other hand, the coefficient for female is positive and significant in all specifications. This means that women are more likely to enroll in honors Math 2 than men. This result, along with the analogous conclusion for English in the next subsection, confirms a pattern already documented in the literature (Shettle *et al.*, 2007; Nord *et al.*, 2011; Long *et al.*, 2012): women are more likely to take advanced courses during high school. The female coefficient can be understood as the gender gap in the probability of enrolling in honors Math 2 for proficient students. Hence, a negative coefficient for being a non-proficient woman implies that although women are more likely to enroll in honors Math 2, the likelihood is not the same for non-proficient women. In other words, this coefficient represents the gender difference in the reaction to low performance and it implies that non-proficient women are less likely to enroll in honors Math 2, but the same is not true for non-proficient men.

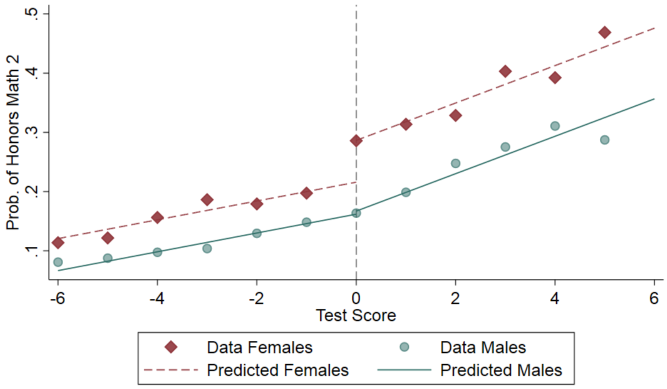
Table 1.4: Probability of Taking Honors Math 2. ±6 bandwidth

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Female (F)	0.119***	0.108***	0.108***	0.108***	0.107***	0.108***	0.108***	0.110***
	(0.014)	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)	(0.011)	(0.011)
Non-Prof.	-0.005	-0.006	-0.006	-0.006	-0.006	-0.007	-0.007	-0.008
	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.005)	(0.005)	(0.006)
F*Non-Prof.	-0.065***	-0.054***	-0.053***	-0.052***	-0.052***	-0.049***	-0.049***	-0.050***
	(0.016)	(0.012)	(0.012)	(0.012)	(0.012)	(0.011)	(0.012)	(0.011)
Math Test MS						0.101***	0.100***	
						(0.009)	(0.008)	
Reading Test MS							0.003	
							(0.006)	
(Math-Reading) Test MS								0.028***
								(0.005)
School FE		✓	✓	✓	✓	✓	✓	✓
Year FE			✓	✓	✓	✓	✓	✓
EDS				✓	✓	✓	✓	✓
Race					✓	✓	✓	✓
Mean	0.23	0.23	0.23	0.23	0.23	0.23	0.23	0.23
R^2	0.06	0.22	0.22	0.23	0.23	0.25	0.25	0.23
N	14,918	14,918	14,918	14,918	14,918	14,184	14,146	14,146

Note: The dependent variable is the same across all specifications: one when taking honors Math 2, zero otherwise. Each specification includes indicator variables for being a female (F), for being non-proficient (Non-Prof) on the Math 1 EOC test, and the interaction of those two variables. Additionally all specifications include a first-degree polynomial of the running variable with different slopes above and below the proficiency threshold. Reading Test MS: standardized middle school reading test score. Math test MS: standardized middle school math test score. (Math-Reading) Test MS: difference between the standardized math and reading middle school test scores. School FE: school fixed effects. EDS: Economically Disadvantaged Student. Standard errors are clustered at test score level.*Significant at 10%, **5%, ***1%.

A visual representation of this results is shown in Figure 1.2.[15] The proficiency cutoff is represented by zero. To the left of zero students are considered non-proficient. There is no discontinuity at the cutoff for men (green solid line). However there is a discontinuity at the cutoff for women (maroon dashed line), non-proficient women are less likely enroll in honors Math 2 than proficient ones.

Figure 1.2: Regression Discontinuity, Math



Note: Test scores are normalized such that a score of 0 or more means proficiency. Lines are fitted values from regressing a dependent variable that equals one when taking honors Math 2 on indicators for being female and non-proficient on the Math 1 EOC test, the interaction of those two variables, and a first-degree polynomial of the running variable with different slopes above and below the proficiency threshold, within ±6 of the cutoff. Standard errors are clustered at test score level. The markers represent the proportion of students by gender that take honors Math 2 within each 1 point bin. *Significant at 10%, **5%, ***1%.

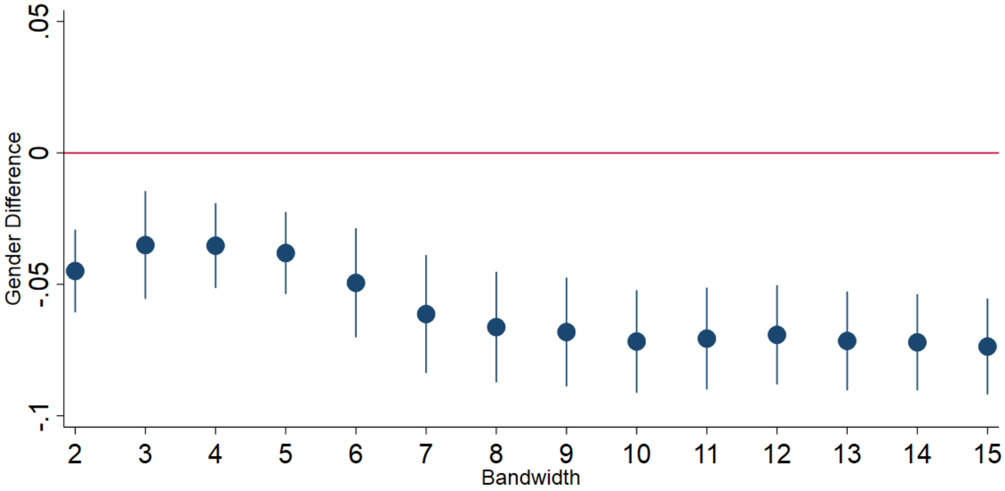
In Table 1.4, the coefficient on the interaction between being a woman and receiving a non-proficient score on the test varies between -6.5 and -4.9 percentage points depending on the specification. High school fixed effects (FE) account for differences across high schools that can affect the likelihood of enrolling in honors classes (for example, high school culture about promoting the advanced classes) and they reduce the gender difference in the discontinuity by about one percentage point. Adding year fixed effects and controlling for race and EDS does not change the effect of being a non-proficient women a considerable amount, which is expected given that these

[15] This is the graphical representation of the specification in column (1) in Table 1.4.

variables do not show a discontinuity at the threshold (See Table 1.2).

Columns (6)-(8) in Table 1.4 control for middle school test scores in math and reading. The objective is to control for the ability of the students in these two subjects right before starting high school, and possible comparative advantages. The literature suggests that women have a comparative advantage in verbal skills, while men have comparative advantage in math (Aucejo and James, 2021; Breda and Napp, 2019; Delaney and Devereux, 2019). However, the coefficient of interest does not change in a meaningful way when past test scores are controlled for, each one individually or when the difference between the two is accounted for. Nonetheless, the middle school math test score seems to be relevant in explaining the regular versus honors decision for Math 2, the higher the score the more likely the student is to enroll in the honors class.

Figure 1.3: Estimates of the Difference in the Discontinuity for Different Bandwidths, Math



Note: Spikes represent 90% confidence intervals. Dots are the coefficients for being a non-proficient female in the preferred specification: an indicator for taking honors Math 2 regressed on indicators for female and being non-proficient on the Math 1 EOC test, the interaction of those two variables, a first-degree polynomial of the running variable with different slopes above and below the proficiency threshold, year and high school FE, indicators for EDS, race and controls for middle school test scores. Standard errors are clustered at test score level. Optimal bandwidth is between 6 and 7.

Results in Table 1.4 are qualitatively robust to different bandwidths.[16] The preferred specification controls for high school and year FE, EDS, race, and middle school test scores (column (7) in Table 1.4). Figure 1.3 shows the estimates of the gender difference in the reaction to non-proficiency for different bandwidths around the cutoff, for this specification.[17] Regardless of the bandwidth chosen, the estimates of the gender difference are consistently between 4 and 7 percentage points, and statistically significant in all cases. This means that women react significantly more strongly to low performance on the Math 1 EOC test than men. Comparing these estimates to the proportion of women that score within 6 points of the cutoff and take honors Math 2 (27%), roughly translates the effect of getting a non-proficient score to a 18 to 24% reduction in the likelihood of enrolling in the honors version of Math 2.

1.4.2 English

Table 1.5 presents the results of the estimation of model (1.1) for English within a bandwidth of ±5 from the cutoff. The optimal bandwidth is between 4 and 5, according to the procedures suggested by Imbens and Kalyanaraman (2012) and Calonico *et al.* (2019).[18] The outcome variable is equal to one when the student enrolls in the advanced version of English 3, zero otherwise. Advanced English 3 includes the honors class and AP or IB classes when offered. As in the math case, the coefficient of interest is the interaction between being a woman and getting a non-proficient score on the English 2 EOC test, (F*Non-Prof). This coefficient is an estimate of the effect of being a non-proficient woman on the probability of enrolling

[16] Appendix Tables A.3 and A.4 show the analogous results for bandwidths ±5 and ±7, respectively.

[17] See Appendix Figures A.4, and A.5 for different specifications.

[18] Results for ±4 and ±6 can be found in Appendix Tables A.6 and A.7, respectively.

in advanced English 3. Hence it represents the estimated gender difference in reaction to low performance.

The difference across the specifications in Table 1.5 are the controls included in each case, which follows the same pattern as in the math results. The sample size reduces when controlling for final grades and middle school test scores for the same reason as in math, it was not possible to recover them for all the students in the sample. Nonetheless, the proportion of students choosing the advanced version of English 3 is constant across all specifications, around 42%.

Regardless of the specification, the coefficient for being non-proficient is always close to zero and not statistically significant. This suggests that the probability of men enrolling in advanced English 3 does not change because of a non-proficient result on the English 2 test. On the other hand, the coefficient for female is always positive, which suggests women are more likely to enroll in advanced English classes than men. However, the coefficient for the interaction between being a woman and getting a non-proficient score on the English 2 EOC test is not always statistically significant, which suggests that in the case of English women and men might not be reacting differently to a non-proficient score. This result is illustrated in Figure 1.4 where there is no discontinuity at the cutoff (zero) for both genders.[19] Once high school fixed effects are included (column (2)) precision improves, which makes the gender difference of about 2 percentage points significant (p-val. < 0.1). This effect does not change in a meaningful way when controlling for year, race and EDS, which is expected given the results from Table 1.3. The inclusion of middle school test scores, each one individually or the difference between the two, does not change the point estimate much, but the significance changes when controlling for both tests separately.

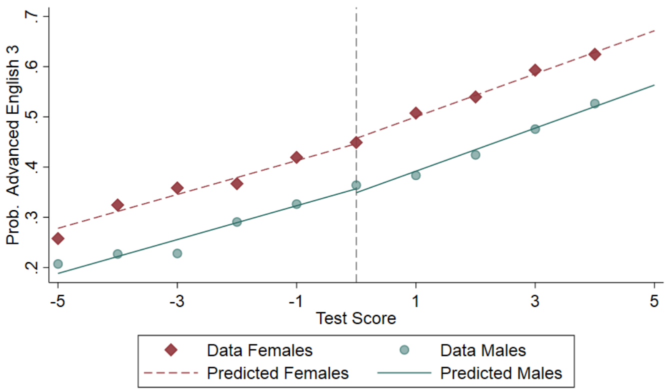
[19] Figure 1.4 is the graphical representation of column (1) Table 1.5

Table 1.5: Probability of Taking Advanced English 3. ±5 bandwidth

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Female (F)	0.108***	0.097***	0.097***	0.103***	0.104***	0.118***	0.134***	0.111***
	(0.006)	(0.006)	(0.006)	(0.005)	(0.005)	(0.006)	(0.005)	(0.005)
Non-Prof.	0.008	0.011	0.011	0.009	0.009	0.009	0.011	0.010
	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.009)	(0.007)	(0.009)
F*Non-Prof.	-0.018	-0.022*	-0.022*	-0.020*	-0.020*	-0.024*	-0.021	-0.023
	(0.012)	(0.011)	(0.011)	(0.010)	(0.010)	(0.012)	(0.012)	(0.013)
Reading Test MS						0.119***	0.070***	
						(0.006)	(0.004)	
Math Test MS							0.128***	
							(0.007)	
(Math-Reading) Test MS								0.054***
								(0.006)
School FE		✓	✓	✓	✓	✓	✓	✓
Year FE			✓	✓	✓	✓	✓	✓
EDS				✓	✓	✓	✓	✓
Race					✓	✓	✓	✓
Mean	0.42	0.42	0.42	0.42	0.42	0.42	0.42	0.42
R^2	0.06	0.21	0.21	0.23	0.23	0.25	0.28	0.24
N	27,072	27,072	27,072	27,072	27,072	25,175	25,137	25,137

Note: The dependent variable is the same across all specifications: one when taking an advanced English 3 class, zero otherwise. Each specification includes indicator variables for being a female (F), for being non-proficient (Non-Prof) on the English 2 EOC test; and the interaction of those two variables. Additionally all specifications include a first-degree polynomial of the running variable with different slopes above and below the proficiency threshold. Reading Test MS: standardized middle school reading test score. Math test MS: standardized middle school math test score. (Math-Reading) Test MS: difference between the standardized math and English middle school test scores. School FE: school fixed effects. EDS: Economically Disadvantaged Student. Standard errors are clustered at test score level.*Significant at 10%, **5%, ***1%.

Figure 1.4: Regression Discontinuity, English

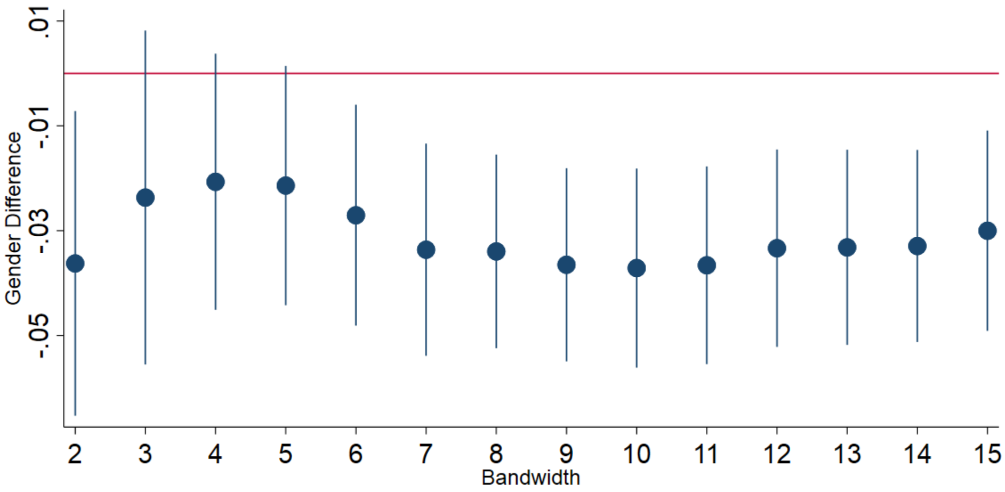


Note: Test scores are normalized such that a score of 0 or more means proficiency. Lines are fitted values from regressing a dependent variable that equals one when taking advanced English 3 on indicators for being female and non-proficient on the English 2 EOC test, the interaction of those two variables, and a first-degree polynomial of the running variable with different slopes above and below the proficiency threshold, within ±5 of the cutoff. Standard errors are clustered at test score level. The markers represent the proportion of students by gender that take advanced English 3 within each 1 point bin. *Significant at 10%, **5%, ***1%.

Figure 1.5 plots the gender difference in the reaction low performance when estimating the preferred specification that controls for high school and year fixed effects, race, EDS and middle school test scores for different bandwidths.[20] Although not always statistically significant, the coefficient for being a non-proficient women is always negative and relatively stable, between 2 and 4 percentage points, across bandwidths which suggest the existence of a small negative effect that is not precisely estimated for the smaller bandwidths. This implies that women also react differently than men when performing poorly on the English 2 EOC test, however the effect seems smaller than in math. Comparing these estimates to the proportion of women who score within 5 points of the cutoff and take advanced English 3 (47%), this roughly translates the effect of women getting a non-proficient score to a 4 to 8.5% reduction in the likelihood of enrolling in the advanced version of English 3, less than the estimated

[20] See Figures A.7, and A.8 for other specifications.

Figure 1.5: Estimates of the Difference in the Discontinuity for Different Bandwidths, English



Note: Spikes represent 90% confidence intervals. Dots are the coefficients for being a non-proficient female in the preferred specification: an indicator for taking advanced English 3 class regressed on indicators for female, being non-proficient on the English 2 EOC test, the interaction of those two variables, a first-degree polynomial of the running variable with different slopes above and below the proficiency threshold, high school and year FE, indicators for EDS, race and middle school test scores. Standard errors are clustered at test score level. Optimal bandwidth is between 4 and 5.

effect for math.[21]

1.4.3 Heterogeneity in the Main Effects

The possibility of heterogeneous effects is studied in Appendix Tables A.5 and A.8 for math and English, respectively. In these tables, the variable of interest, F*Non-Prof., is interacted with various socioeconomic and academic performance variables in order to determine if different reactions to low performance by gender vary across groups.

For the math case, there is no evidence that the different reaction to low perfor-

[21] Figure 1.6 plots together the gender difference in the discontinuity for math and English. It suggests that women react more to non-proficiency in math than in English.

mance by gender differs across economic status, race or high/low academic performance during middle school. In the English case, there is evidence that black non-proficient women are more likely to enroll in advanced English 3 than non-proficient non-black women, however men do not react to low performance regardless of their race. Additionally, economically disadvantaged students seem to react differently to low performance than non-disadvantaged students, but the reaction does not differ across genders.

1.5 End of High School Outcomes

In this section, I discuss the results from estimating model (1.1) when the outcomes are end of high school measures like the number of advanced classes taken during high school or plans to attend a 4-year college, instead of the earlier decisions examined before. These results provide some evidence of the longer term effects of non-proficiency early on during high school, and how these effects are different across genders. Results are only presented for the preferred specification at their respective optimal bandwidths.

1.5.1 Math

Table 1.6 presents the results for different outcomes realized during the last years of high school: the probability of taking at least one higher-level advanced math class, the total number of higher-level advanced math courses taken during high school, and the plans to attend a 4-year college. The set of higher-level advanced math classes includes any senior level honors class, as well as any AP, IB or community college math course.

Women are more likely to take higher-level advanced classes, as the first two columns of Table 1.6 show. On the extensive margin, non-proficient and proficient

Table 1.6: Effect of Non-Proficiency on Math 1 Test on Different Outcomes.

	Prob. higher-level advanced math class	Number of higher-level advanced math classes	4-year College Plan
	(1)	(2)	(3)
Female (F)	0.108***	0.165***	0.141***
	(0.007)	(0.006)	(0.011)
Non-Prof.	0.016	0.045**	-0.002
	(0.010)	(0.016)	(0.015)
F*Non-Prof.	-0.039***	-0.064***	-0.029**
	(0.010)	(0.013)	(0.011)
Mean	0.35	0.47	0.40
R^2	0.27	0.34	0.20
N	13,861	12,004	10,955

Note: Dependent variable indicated at the top of each column. Prob. higher-level advanced math class: one if the student ever took a higher-level advanced math class, zero otherwise. 4-year College Plan: one if student plans to attend a 4-year college, zero otherwise. Each column follows the same specification: dependent variable regressed on a variable equal one when the student is a woman, a variable equal to one when the student is deemed non-proficient on the Math 1 EOC test, the interaction of those two and a first degree polymonial of the test score with flexible slopes above and below the proficiency cutoff. All columns include year and high school fixed effects, controls for EDS, race, and middle school test scores. EDS: Economically Disadvantaged Student. Each column runs a regression withing the optimal bandwidth given the dependent variable: 5, 6 and 4 for columns (1)-(3), respectively. Standard errors are clustered at test score level. *Significant at 10%, **5%, ***1%.

men are equally likely to take a higher-level advanced math class. However, women react differently than men to non-proficiency: the likelihood of ever taking higher-level advanced classes is 4 percentage points lower among non-proficient women.

On the intensive margin, men right below the proficiency cutoff take more higher-level advanced classes, and the opposite is true for women. Although in general women take more higher-level advanced math classes than men, the number is not the same across proficient and non-proficient women, with the latter group taking less higher-level advanced courses. These gender differences in high school preparation could have implications for college major decisions, because women and men might not be equally prepared to undertake certain majors that require a strong mathematical background (Delaney and Devereux, 2019) like STEM majors or Economics.

Finally, women are more likely to plan to attend a 4-year higher education institution. Even though the probability of planning to attend a 4-year college does not change across male students deemed proficient or non-proficient on the Math 1 EOC test, non-proficient women are less likely to state they will attend a 4-year college than proficient ones. This means that the effect of non-proficiency on a first-year test has different effects across gender even at the end of high school on outcomes as important as college attendance.

1.5.2 English

Table 1.7 shows the results for some end of high school outcomes of interest: the likelihood of taking higher-level advanced English classes during high school, the total number of higher-level advanced English classes taken during high school, and plans to attend a 4-year college. The set of higher-level advanced English classes includes any junior or senior honors English class, as well as any AP, IB or community college English course. This set of courses includes more than just the courses considered as possible options for English 3 in the main results.

Women are more likely to take higher-level advanced classes regardless of the subject as shown in the first two columns of Table 1.7. Similar to higher-level advanced math, on the extensive margin non-proficient and proficient men are equally likely to take a higher-level advanced English class. Given that the coefficient for being a non-proficient women is not statistically different from zero, women do not react differently than men to getting a non-proficient English 2 test score. On the intensive margin, the number of higher-level advanced English classes taken by non-proficient and proficient men do not differ at the cutoff. However, non-proficient women tend

Table 1.7: Effect of Non-Proficiency on English 2 Test on Different Outcomes.

	Prob. of higher-level advanced English class	Number higher-level advanced English classes	4-year College Plan
	(1)	(2)	(3)
Female (F)	0.140***	0.314***	0.084***
	(0.008)	(0.011)	(0.005)
Non-Prof.	0.006	0.014	0.018
	(0.013)	(0.022)	(0.010)
F*Non-Prof.	-0.012	-0.051**	-0.023*
	(0.014)	(0.022)	(0.012)
Mean	0.50	0.92	0.38
R^2	0.27	0.33	0.20
N	24,683	24,683	28,701

Note: Dependent variable indicated at the top of each column. Prob. of higher-level advanced English class: one if the student ever took a higher-level advanced English class, zero otherwise. 4-year College Plan: one if student plans to attend a 4-year college, zero otherwise. Each column follows the same specification: dependent variable regressed on a variable equal one when the student is a woman, a variable equal to one when the student is deemed non-proficient on the English 2 EOC test, the interaction of those two and a first degree polymonial of the test score with flexible slopes above and below the proficiency cutoff. All columns include controls for school and year fixed effects, EDS, race, and middle school test scores. EDS: Economically Disadvantaged Student. Each column runs a regression withing the optimal bandwidth given the dependent variable: 5, 5 and 6 for columns (1)-(3), respectively. Standard errors are clustered at test score level. *Significant at 10%, **5%, ***1%.

to take a lower number of higher-level advanced English classes than the proficient ones.

As in the math case, women are more likely to state plans to attend a 4-year college. There is no difference in the likelihood of planning to attend a 4-year college between proficient and non-proficient male students. However, women react differently than men when they get a non-proficient score on the English test. Non-proficient women are 2.5 percentage points less likely to state plans of attending a 4-year college than proficient ones. This means that the effect of non-proficiency on a test early during high school has different effects across gender even at the end of high school on an outcomes as important as college attendance, regardless of the subject.

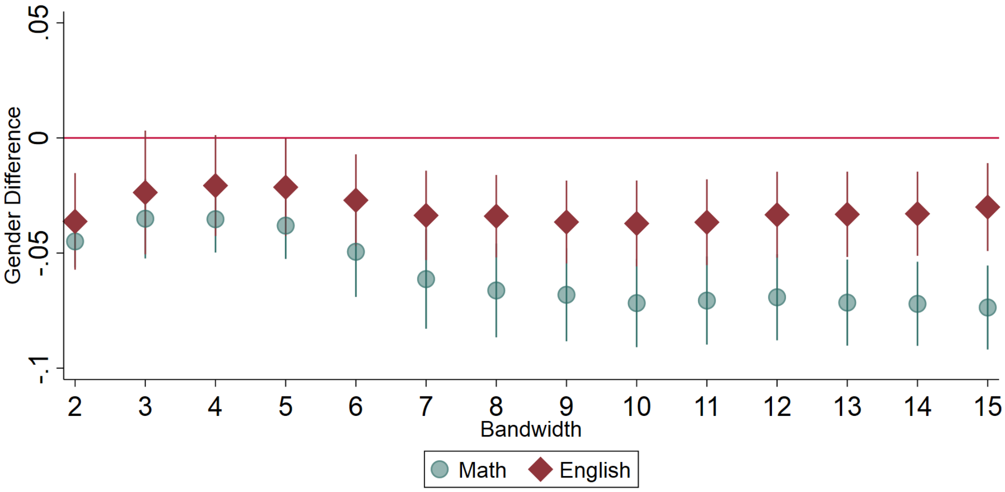
1.6 Possible Mechanisms

The results from the empirical section only provide evidence about the existence of a gender difference in the reaction to low performance, but they do not give any information about the mechanisms driving the results. However, my findings are consistent with the idea that women interpret their non-proficiency status as a discouraging signal of their ability and this affects their future behavior in terms of class enrollment and college attendance.

Experimental evidence suggests that women are more likely to attribute negative feedback to ability while men tend to attribute it to bad luck (Roberts and Nolem-Hoeksema, 1989; Shastry *et al.*, 2020). Women tend to update their beliefs more pessimistically than men after receiving negative feedback, even when controlling for performance (Berlin and Dargnies, 2016). Moreover, there is experimental evidence that women are not only less willing to compete, but also less likely to do it after losing (Buser and Yuan, 2019a), which is consistent with women not taking "harder" classes after performing poorly on the standardized tests.

Additionally, given that the effects of low-performance are stronger in math than in English (see Figure 1.6), it seems reasonable that the mechanisms are related to the stereotype associated with the domain. For instance, using a lab experiment, Kiefer and Shih (2006) concludes that women are more sensitive to feedback on tests that measure math ability than verbal ability and the opposite pattern is true for men. Similarly, Coffman *et al.* (2019) finds that male and female beliefs about performance on a task respond more to feedback when the task is gender congruent than when it is not, i.e. men's beliefs respond more to feedback on male-related tasks and women respond more when the feedback is about female-related tasks. In a college context,

Figure 1.6: Estimates of the Difference in the Discontinuity for Different Bandwidths



Note: Spikes represent 90% confidence intervals. Dots are the coefficients for being a non-proficient female in the preferred specification: the indicator for taking advanced English 3 or honors Math 2 class regressed on indicators for female, being non-proficient on the English 2 or Math 1 EOC test, the interaction of those two variables, a first-degree polynomial of the running variable with different slopes above and below the proficiency threshold, high school and year FE, indicators for EDS, race, and middle school test scores. Standard errors are clustered at test score level.

Kaganovich *et al.* (2021) finds that female grade sensitivity is category-specific, women are more sensitive to low grades in STEM and Business majors (male-dominated areas) than in Social Science majors. Also, Ellis *et al.* (2016) suggest that the high dropout rate of women from STEM majors can be related with women's lack of confidence in their math ability, rather than with actual lack of ability.

Another possibility that cannot be ruled out is that parents and/or teachers react differently to a negative outcome depending on the student's gender, which affects students' enrollment decisions. For example, in an experiment with Norwegian families, Tungodden and Willén (2022) document that parents of boys chose a competitive task for their sons more often than parents of girls. However, this explanation is not as convincing because there is evidence of similar gender differences in behavior after instances of "low performance" in college settings, where the influence of parents on the students decisions is lower (Rask and Tiefenthaler, 2008; Kaganovich *et al.*, 2021;

Kugler *et al.*, 2021). Regarding teachers, Gentrup and Rjosk (2018) finds that teachers have higher math achievement expectations for boys than for girls, but higher reading achievement expectations for girls. Additionally, girls' math achievement is more negatively affected by low teacher expectations and benefits less from high expectations than boys' achievement. Lavy and Sand (2018) find that primary school teachers' biased expectations in favor of boys encourage them to enroll in advanced math and science courses during high school, but discourage girls.

However, all these explanations can interact with each other, since parents and teachers expectations/stereotypes can affect students self-confidence and reaction to setbacks, impacting their behavior differently across areas. Therefore, there are several possible mechanisms that could explain the gender difference in reaction to low performance that I find, and further research is required to determine how relevant they are and how they interact with each other.

1.7 Conclusions

This chapter uses a RD design to study the decision of North Carolina public high school students to enroll in advanced math or English classes after learning about their performance on EOC tests. The objective is to determine whether there is a gender difference in the reaction to low performance, and if the reaction depends on the subject. Men are usually associated with math and science while women are associated with languages and literature. Therefore, the idea is to determine if the reaction to low performance varies across subjects with different gender stereotypes.

I find a gender difference in the reaction to low performance. In particular, although women are more likely to enroll in advanced classes, the probability of doing so changes depending on their test score performance, while men's likelihood of taking advanced classes does not change with their proficiency status. I find roughly a reduction of 18 to 24% in the probability of enrolling in honors Math 2 when a woman is not proficient on the Math 1 EOC test, but no reduction for low-performance men. The effect seems to be smaller in English than in math, with an estimated reduction of about 4 to 8.5% in the probability of enrolling in advanced English 3 among women that preformed poorly on the English 2 EOC test, and once again no effect of low performance for men. These results suggest that the reaction to low performance might depend on the area in which the hardship is experienced.

If these findings are extrapolated to a career setting, they could be a plausible explanation for women being under-represented in high-level positions. If women and men react to setbacks in different ways, an early defeat might preclude a women from advancing, especially in a male-dominated area. Although the methodology used here does not provide evidence about the mechanisms driving the results, the literature points towards women interpreting their non-proficiency status as a discouraging signal of their ability which affects their future behavior in terms of class enrollment and college attendance. However, further research is required to investigate this or other mechanisms driving the gender differences in the reaction to low performance.

Chapter 2

GENDER, GRADE SENSITIVITY, AND MAJOR CHOICE

2.1 Introduction

It has been documented that women in STEM and other male-dominated areas like Economics are more sensitive to grades than men, in the sense that the probability of women continuing their studies in or switching out of those fields is more affected by their performance in relevant courses at the beginning of their college career (Rask and Tiefenthaler, 2008; Ost, 2010; Goldin, 2015; Kugler *et al.*, 2021). There is significant interest from universities, governments, and policymakers around the world in closing the gender gap in these areas.[1] In order to design policies that effectively encourage the participation of women in traditionally male-dominated fields, it is crucial to identify the primary factors that explain these gender differences in behavior. However, the mechanisms driving these grade sensitivity differences are still poorly understood. Therefore, this project studies why women and men react differently to grades during college and how this behavior impacts their decision to persist or switch out of a given major.

Understanding why talented women with the potential to succeed in male-dominated fields drop out because of less-than-stellar grades in an introductory class is important for several reasons. From a gender equality perspective and given that field of study is a key determinant of occupational choices and earnings (Gemici and Wiswall,

[1] For example, in October 2021, the White House released the National Strategy on Gender Equity and Equality which “seeks to close gender gaps in STEM fields so that women and girls can shape the workforce of the future.”

2014; Golan and Sanders, 2019; Patnaik *et al.*, 2021), this could have important implications for the labor market outcomes of highly skilled women because jobs in male-dominated fields like STEM, Economics, and business pay higher wages than other areas (Altonji *et al.*, 2012, 2014, 2016). Understanding these mechanisms is also important for a society interested in promoting economic growth through the most efficient possible allocation of talent and resources across fields of study and occupations since a higher rate of women dropping out of traditionally male majors is potentially consistent with a misallocation of talent and labor market inefficiencies (Hunt, 2016; Hammond *et al.*, 2020).

I document the grade sensitivity patterns among undergraduate students at Arizona State University (ASU), one of the largest public universities in the United States. ASU's administrative data allows me to trace the trajectory of students as they progress through their college careers, including all fields of study switches. Using a logit model, I calculate the probability that freshmen students remain in their first-year major conditional on their first-year GPA. Majors are grouped into three broad categories: STEM, Business/Economics (BEC), and Humanities/Social Sciences (SSH). I find that the gender gap (male-female) in the probability of staying in STEM and BEC majors increases as GPA decreases. However, such a relationship is not observed in other majors like SSH, where the gender gap in the probability of staying in those majors remains constant regardless of first-year GPA.

The fact that women's probability of persisting in STEM and BEC majors is more responsive to their first-year GPA than men's suggests that women care more about grades than men. However, it is not clear why those gender differences in sensitivity to grades arise, and what exactly are the mechanisms through which they impact major decisions. Administrative data provide information about students' actual choices. Therefore, concerns about selection due to unobserved tastes for each major limit the ability of these data to shed light on what exactly leads to those patterns. Given this limitation, I designed an online survey to collect novel data that allow me to (1) quantify students' sensitivity for grades, and (2) investigate how gender differences in grade sensitivity impact students' decision to persist or switch out of a given major. Around 2,000 ASU students participated in the study.

To quantify the gender differences in grade sensitivity, I use hypothetical choice scenarios. This methodology has been used in a wide variety of contexts, for example, to study preferences for reliable electricity services (Blass *et al.*, 2010), political candidates (Delavande and Manski, 2015), workplace attributes including valuation of harassment risks at work (Wiswall and Zafar, 2018; Folke and Rickne, 2022), and neighborhood characteristics (Koşar *et al.*, 2022) among others. This approach allows me to collect data on students' preferences for different attributes that characterize majors: average GPA at graduation, average weekly study time, and average earnings at a full-time job after graduation. The survey includes 10 different individual-specific hypothetical scenarios. In each scenario, participants report the probability that they would choose each of the three majors (SSH, BEC, STEM) given the attributes in that scenario. This design generates a panel of probability choices, which allows me to estimate preferences at the individual level.

I find that on average, students have preferences reflecting a distaste for study time, and a taste for a higher GPA at graduation. Based on the estimated preferences for average GPA at graduation, I calculate a willingness-to-pay (WTP) measure. This WTP measure is interpreted as the amount of annual earnings a participant is willing to forego for a one-point increase in the average GPA at graduation in a given major. I find that on average, students are willing to pay 16% ($8,309) of annual earnings for an extra average GPA point at graduation. Conditional on background characteristics, I find that women are willing to pay $3,057 more of the annual earnings relative to men.

I interpret this difference as the gender gap in grade sensitivity. Moreover, splitting the sample by students' reported major of enrollment suggests that the gender gap is concentrated among STEM/BEC students, for whom the gender difference in WTP for a GPA increase reaches $3,760. This result is consistent with the results from the administrative data: women in STEM and BEC majors are more sensitive to grades than men, but this gap is not observed in other majors.

These results are also consistent with the literature on grade sensitivity, although there are several differences in my methodology relative to the previous studies. For example, some of the existing work focuses on students from specific majors instead of looking at several disciplines as I do. In Ost (2010), the authors study the persistence of students that intend to major in science. Using administrative data, they exploit the variation that comes from students switching out of science majors to find that women majoring in physical sciences are more responsive to grades than men: an increase in the GPA from physical sciences courses increases the probability of persistence in the major more for women than for men. Goldin (2015) uses the same type of variation but focuses on students that take economics introductory classes. She documents that women that receive grades lower than B in their introductory economics courses are less likely to graduate with an economics major than similarly achieving men. Also looking only at students that took an introductory economics class, Rask and Tiefenthaler (2008) shows that when deciding whether to take one more economics course, women are significantly more responsive to the grades they received in previous economics classes than men, especially women in the bottom half of the grade distribution.

On the other hand, Kugler *et al.* (2021); Kaganovich *et al.* (2021), and Ahn *et al.* (2022) use detailed administrative data across several disciplines to study the effect of grades on major choices. In Ahn *et al.* (2022), the authors estimate a structural model of course choices and grading policies and find that women value grades more than men. The variation in the data that they exploit to identify the value of grades comes from individuals sorting into different majors based on where their abilities are rewarded more. Similarly, Kugler *et al.* (2021) exploits the variation that comes from students changing majors during college in a logit regression to conclude that women are more likely to switch out of male-dominated STEM majors if they have a low GPA. Using similar variation in their administrative data and a multinomial logit model, Kaganovich *et al.* (2021) finds that the probability that women persist in STEM majors is more responsive to the grades they receive in that major relative to men. A key difference between my work and the existing literature is that my approach is based on stated preferences instead of revealed preferences from administrative data. In my case, the preferences for grades are identified from within-individual variation in stated choices, and the estimation is carried out separately for each individual. This reduces concerns about selection and allows me to estimate the complete distribution of preferences and WTP for GPA without imposing any parametric restrictions. Additionally, this approach allows me to quantify the sensitivity to grades in an easily interpretable way as a measure of willingness-to-pay in dollars.

Also, most of the work on grade sensitivity remains agnostic about the mechanisms driving the gender differences in reaction to grades.[2] Therefore, I further contribute to the literature by collecting and analyzing information about different hypotheses that could explain the sensitivity patterns observed in the data. Many potential mechanisms could drive these differences. For example, gender differences in risk aversion (De Paola and Gioia, 2012), willingness to compete (Buser *et al.*, 2014), self-confidence (Ellis *et al.*, 2016; Moakler and Kim, 2014), and beliefs about what it takes

[2] One exception is Kaganovich *et al.* (2021) which finds that tastes for different majors are important to understand the gender differences in grade sensitivity.

to graduate from a male-dominated major (Owen, 2020). I collect information about some of these hypotheses and discuss them later. However, a much less explored possibility and the primary focus of this paper is anticipated gender discrimination in the labor market (Steele *et al.*, 2002; Alston, 2019).

There is evidence that women face gender discrimination in the labor market, especially in the form of different or more rigorous standards than men in terms of hiring and promotion decisions. In a lab experiment, Foschi *et al.* (1994) finds that men exhibit a double standard in their hiring decisions for engineering positions. When the male candidate has a better performance than the female one, the male candidate was chosen more and was considered more competent and suitable for the job. However, the same was not true when the candidate with the best performance was a female. In Goldin and Rouse (2000), the authors provide evidence of sex-biased hiring in symphony orchestras against women, since a blind audition that conceals the candidate's identity (and gender) increases the probability of women being hired. Quintero (2008) finds that during the recruitment process for government jobs in Spain, women are treated worse than men even when there is no evidence of lack of ability, and men are subject to a more lenient standard. In Funk and Parker (2018), the authors conduct a nationally representative survey of U.S. adults and find that 50% of the women that work in STEM jobs report having experienced gender discrimination at work.[3] Among the participants that say their gender has made it harder to succeed in their job, 14% say it is because they are held to different standards. In a survey of female scientists, Williams *et al.* (2014) finds that 64% of the participants believe they needed to provide more evidence of competence than others to prove themselves to their colleagues.

Given this evidence, female students might anticipate facing gender discrimination

[3] The figure is 78% for majority-male workplaces.

in the labor market, particularly in male-dominated fields like STEM and business. I develop a theoretical framework to formalize the intuition behind how these beliefs can lead to gender differences in grade sensitivity and major choices. In the model, students are enrolled in a science major but can switch to a humanities major after receiving new information about their ability in the form of grades. I allow the utility of each major to depend on the probability of finding a job. Students believe that employers make their hiring decisions as in the model of labor market discrimination in Coate and Loury (1993), where if employers discriminate against women they set higher or more rigorous standards to hire them. This means that women believe they have lower chances of getting a job in that field. By allowing major choices to be affected in this way by discrimination, and assuming women believe they are discriminated against in the science field, women and men that receive the same grades make different decisions about staying or leaving the major. Women are more likely to leave the science major than men who get the same grades because they believe they will be treated differently in the labor market given their gender.

In the survey, I collect data about perceived gender discrimination in each field and beliefs about the standards faced in the labor market to get a job. For each major, I ask participants how likely they think it will be that finding a job in that field would be harder because of their gender, and how likely it would be that their boss or peers would treat them differently because of their gender. Using their responses, I create an anticipated gender discrimination index for each major. I find that men believe that they are less likely to experience gender discrimination in the labor market than women. Additionally, women believe they are more likely to face gender discrimination in the STEM/BEC labor market than in SSH.

College GPA is commonly used in the hiring process for entry-level positions since a higher GPA is associated with cognitive ability, job performance, and other characteristics that the recruiters consider important for the job (McKinney and Miles, 2009, Toft Hansen *et al.*, 2023). Additionally, there is evidence that a higher GPA increases students' probability of getting a job (McKinney *et al.*, 2003; Quadlin, 2018; Kessler *et al.*, 2019). Therefore, to understand students' beliefs about the labor market standards they will face, I ask them their beliefs about the minimum GPA required to secure a full-time job in each major. On average, participants believe that the standards are lower in the SSH field. Although women anticipate higher standards than men in all fields, they expect to face higher standards in terms of GPA in STEM and BEC majors than in SSH. There is a positive relationship between these beliefs about labor market standards and the beliefs about the likelihood of experiencing gender discrimination in the labor market, particularly for women.

When studying the gender gap in WTP for GPA, I find that the beliefs about GPA standards and anticipated gender discrimination reduce the gap by 48%, making it no longer statistically significant. This means that when comparing men and women that expect to face the same level of gender discrimination and GPA standards in the labor market, on average there is no statistical difference in how much they value grades. In fact, according to an Oaxaca-Blinder decomposition, these beliefs explain 52% of the gender gap in WTP for GPA. These results imply that to understand why women and men value grades differently, especially in STEM and BEC majors, it is important to consider beliefs about labor market standards and gender discrimination.

I do not claim that these beliefs are the only mechanism driving the gender differences in grade sensitivity. However, there are several reasons why anticipated discrimination is an important mechanism worth investigating. First, there is a considerable amount of work on gender discrimination in the labor market, but much less on anticipated discrimination or its relationship with major choices. For example, in economics, Alston (2019) is one of the first papers to study anticipated discrimination as a reason that could explain why women are underrepresented in certain occupations. The author studies the effect of anticipated discrimination on job applicants' decisions to apply or not for a stereotypically male job. On the other hand, I analyze its effect on major choices, which happen earlier in life and impact occupation decisions. In the psychology literature, Steele *et al.* (2002) documents that female undergraduate students in mathematics, science, and engineering majors anticipate encountering more discrimination in their careers than women in the arts, humanities, and social sciences. Therefore, my results contribute to improving our knowledge about the effects of gender discrimination in different aspects of life. Second, some of the other explanations for the gender differences in grade sensitivity rest on inherent differences between women and men like risk aversion, self-confidence, or willingness to compete. However, it is important to investigate mechanisms that instead rest on beliefs about the labor market, like anticipated gender discrimination, because evidence in favor of them suggests very different policy implications. For example, if students' beliefs about gender discrimination are close to the reality of the labor market, then policymakers should aim to solve the discrimination issues in the labor market. Conversely, if students hold inaccurate beliefs, information interventions could be a valuable tool.

I also collect data about self-confidence and beliefs about grades in different fields as potential explanations for why women and men value grades differently. However, they are not able to explain as much of the variation in WTP for GPA as anticipated discrimination.

The rest of the paper is organized as follows. Section 2.2 describes the administrative data and documents gender gaps in grade sensitivity among ASU students. Section 2.3 introduces the survey and describes the sample. Section 2.5 presents the hypothetical scenarios from the survey and section 2.6 explains how to use that data to estimate preferences and WTP for GPA measures. In section 2.7, I focus on anticipated discrimination as a potential mechanism that could explain the gender differences in WTP documented in the previous section, and in section 2.8, I analyze the role of other mechanisms. Finally, section 2.9 concludes.

2.2 Women are More Sensitive to Grades

In this section, I use anonymized transcript-level data for 180,000 first-time freshmen at Arizona State University (ASU), one of the largest public universities in the United States, to provide suggestive evidence that women are more sensitive to grades in STEM and Business majors. The approach in this section is similar to Kaganovich *et al.* (2021).

The administrative data set goes back to the year 2000 and traces the trajectory of students as they progress through their college careers, including all fields of study switches. Majors are grouped into three broad categories: STEM, Business/Economics (BEC), and Humanities/Social Sciences (SSH).[4] I refer to these categories simply as majors.

The probability that freshmen remain in their first-year major conditional on their first-year GPA is calculated from the logit estimation of model (2.1) for each major separately.[5]

$$\mathbb{1}(\text{Stay})_{ikt} = \delta_0 + \delta_1 Female_i + \delta_2 GPA_{ik} + \delta_3 GPA_{i-k} + \mathbf{M}_i + \mathbf{N}_i + \gamma_t + \epsilon_{ik} \qquad (2.1)$$

[4] The SSH category includes any majors that could not be classified as STEM or Business/Economics.

[5] The sample for this exercise consists of students that stay enrolled in college at least until the end of their sophomore year. In other words, it does not include people that dropout at the end of their freshman year. However, the gender differences in the probability of persisting in a given major are robust to including dropouts.

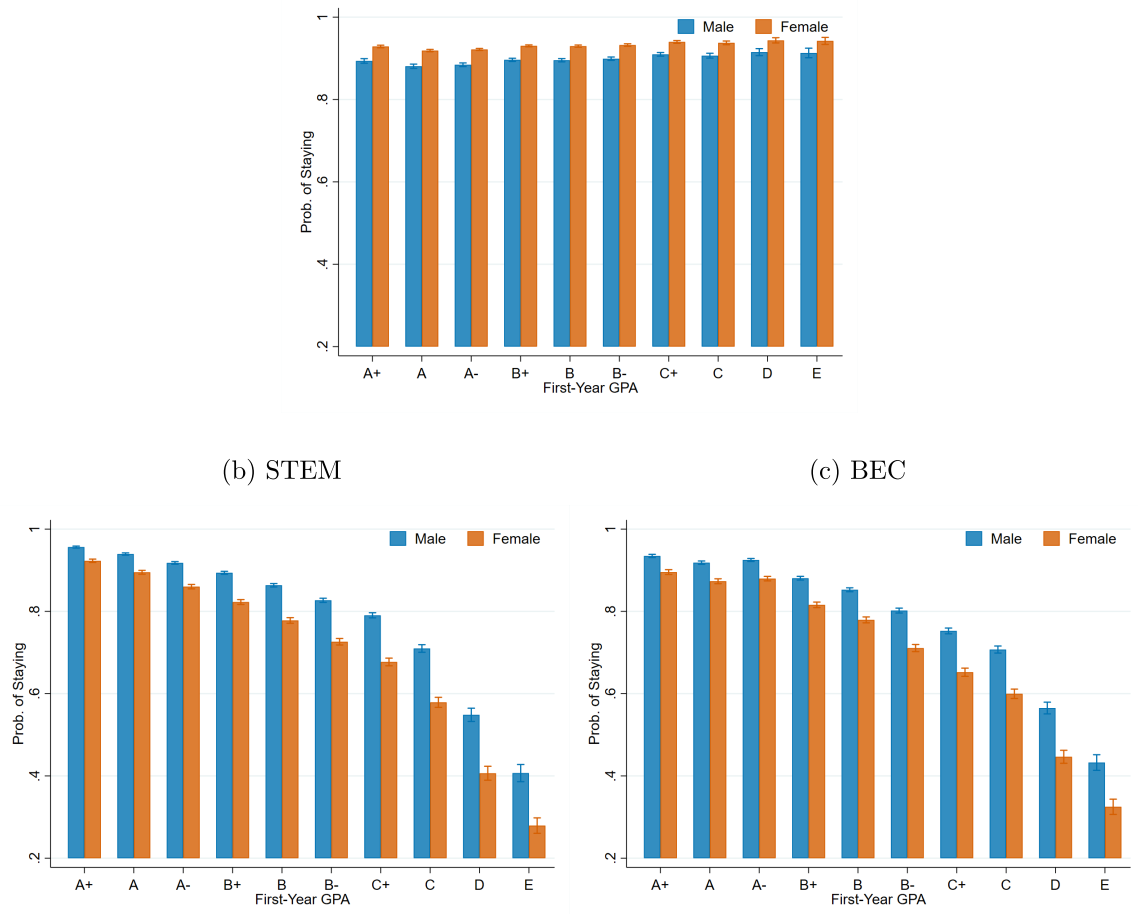
where $k \in \{\text{SSH, BEC, STEM}\}$, $\mathbb{1}(\text{Stay})_{ikt}$ is an indicator variable equal to one when student i from cohort t registered in major k during their freshman year remains in major k during their sophomore year. $Female_i$ is equal to one when student i is female. GPA_{ik} represents cumulative GPA for student i at the end of their freshman year in major k, and GPA_{i-k} is a vector that contains the cumulative GPA in the other majors besides k. To create GPA_{ik} and GPA_{i-k}, all courses were classified into one of the three major categories (SSH, BEC, STEM) and the respective GPA was calculated using only the courses that correspond to that major. $\mathbf{M}_i$ is a set of academic controls: ACT/SAT test scores, high school GPA, and indicators for honors and exploratory students.[6] $\mathbf{N}_i$ includes controls for minority, income, in-state student, and first-generation status. Finally, γ_t represents cohort fixed effects.

The results from this exercise are summarized in Figure 2.1. The bars represent the probability of staying in the major indicated at the top of each panel given the first-year GPA level on the horizontal axis. In panels (2.1b) and (2.1c), the probability of staying in STEM and BEC majors decreases as the GPA decreases, which means that students are more likely to switch out of these majors when they have low grades. Additionally, this pattern is sharper for women than for men, which illustrates the fact that women are more responsive to grades in these majors than men.[7] However, such a gender difference is not observed in SSH in panel (2.1a), where the gender gap in the probability of staying in that major remains constant regardless of first-year

[6] The exploratory indicator identifies students that did not declare a major in their freshman year. However, exploratory students are enrolled in special programs that allow them to explore several majors within an area, which facilitates their classification in one of the three broad categories. The most common exploratory programs are health and life sciences; humanities, fine arts and design; mathematics, technology, engineering, and physics; and social and behavioral sciences.

[7] The difference between the blue and orange bars is statistically different from zero at 1% for all GPA levels.

Figure 2.1: Probability of Persisting in a Major by First Year GPA


(a) SSH

(b) STEM

(c) BEC

Note: Bars represent the probability of staying in the major indicated at the top of each panel given the first-year GPA level on the horizontal axis, estimated from a logit model that regresses an indicator for staying in the same major as in the first year on a female indicator, the GPA in that major, and the GPA in the other majors. All regressions control for minority status, family income, first generation, in-state, honors and exploratory status, ACT/SAT, high school GPA, and cohort FE. Spikes represent 95% CI.

GPA.

These results are consistent with previous literature on grade sensitivity (Rask and Tiefenthaler, 2008; Ost, 2010; Goldin, 2015; Kugler *et al.*, 2021; Kaganovich *et al.*, 2021), and suggest that women care about grades more than men, particularly in STEM/BEC majors. However, due to selection concerns and confounders like tastes for different majors, observational data alone have a limited ability to shed light on what exactly leads to these patterns. When I see people changing majors in the administrative data, it is impossible to know exactly why they are doing it and what is the role of grades in such decisions. For that reason, I designed a survey experiment that allows me to quantify student's sensitivity to grades in a cleaner way, and understand better why women and men could value grades differently and how those differences impact their decision to persist or switch out of a given major. I describe the survey in the next section. Given the similar patterns for STEM and BEC in Figure 2.1, for most of the analysis these two categories will be pooled into one.

2.3 Survey Data

2.3.1 Survey

The data come from an original online survey of undergraduate students at ASU. Students were directly invited to participate via email. Additionally, the study was advertised on the My ASU website, accessible only through the student's ASU ID and password. Students were invited to participate in a study about how they chose their major and the relationship between study time and grades, for which they would enter a lottery for one of 350 $20 eGift Cards. Data collection started on April 5th, 2021 and lasted for about two weeks.

The survey was programmed in Qualtrics. It also collected data on students' demographics, family background, major, academic performance, and study time. The survey instrument can be found here.

2.3.2 Sample

A total of 2,036 respondents completed the survey. 3% of participants that identify as non-binary or decided not to disclose their gender were excluded from the analysis. Additionally, responses in the 1st and 99th percentile of survey duration were excluded, leading to a final sample size of 1,936. The median completion time was 23 minutes (43 minutes on average).

Women comprise 64% of the sample. Although they are over-represented in the survey sample relative to ASU's student population (51% female), there is no differential selection on observables across genders (see Table D.2). This suggests that, in terms of gender differences in background characteristics, the sample is a reasonable representation of ASU students.

For the survey, majors were grouped into the same three broad categories: STEM, Business/Economics (BEC), and Humanities/Social Sciences (SSH).[8] I refer to these categories simply as majors. The last three rows in Table D.2 show the proportion of women and men in each major. The sample includes fewer men in BEC and fewer students in SSH than ASU's student population. However, the gender gap in STEM is the same in the survey sample and the ASU student body (20% gap).

2.4 Major Attributes

As discussed in section 2.2, there is a relationship between students' grades and their persistence in certain majors. Therefore, in the survey, I asked participants

[8] The SSH category includes any majors that could not be classified as STEM or BEC.

Table 2.1: Sample Compared to ASU Population

	Survey			ASU			P-value[c]
	Female	Male	Diff.	Female	Male	Diff.	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Black	0.05	0.03	0.02	0.04	0.03	0.01	0.134
White	0.66	0.70	-0.04	0.46	0.48	-0.02	0.498
Hispanic	0.23	0.18	0.05	0.29	0.23	0.07	0.284
First Generation[a]	0.29	0.23	0.06	0.31	0.23	0.08	0.263
Family Income[b]	102	109	-7.1	126	151	-26	0.181
Freshman	0.22	0.20	0.02	0.26	0.25	0.01	0.776
Sophomore	0.24	0.23	0.00	0.26	0.25	0.01	0.853
Junior	0.30	0.30	0.01	0.22	0.22	0.00	0.806
Senior	0.24	0.27	-0.03	0.26	0.28	-0.02	0.742
ACT	27.71	28.56	-0.85	23.98	25.62	-1.64	0.003
STEM	0.38	0.58	-0.20	0.25	0.46	-0.20	0.689
BEC	0.18	0.21	-0.03	0.18	0.27	-0.10	0.000
SSH	0.44	0.22	0.22	0.57	0.27	0.30	0.001
Sample Size	1,236	700		22,755	21,637		0.000[d]

Note: ASU data includes everyone taking at least one class for credit during the Spring semester of 2021 and attending ASU as their first full-time university. Income and first generation variables for the ASU data are constructed with the first year of available data, which it is not the freshman year all the sample.
[a] Students with no parent with a college degree.
[b] Family income in thousands of dollars.
[c] P-value for whether the gender differences in the survey sample and the ASU population are different.
[d] P-value for the difference in females proportion between the survey sample and ASU population.

to report their beliefs about certain major characteristics including average GPA. In particular, they provided their beliefs about three attributes: average GPA at graduation, average weekly study time, and average earnings at a full-time job after graduation.

Table 2.2 reports the mean and standard deviation of participants' beliefs about each of these attributes for each major by gender. Participants believe that SSH has the highest average GPA at graduation relative to the other two majors. On

Table 2.2: Beliefs about Major Attributes by Gender

	Av. GPA			Av. Study Time			Av. Earnings		
	Female	Male	P-value	Female	Male	P-value	Female	Male	P-value
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
SSH	3.47	3.38	0.000	14.60	12.68	0.000	41.60	40.01	0.005
	(0.27)	(0.30)		(8.26)	(7.49)		(12.56)	(10.74)	
BEC	3.37	3.29	0.000	14.06	13.20	0.024	55.02	53.57	0.067
	(0.31)	(0.31)		(8.17)	(7.89)		(17.66)	(14.80)	
STEM	3.37	3.21	0.000	22.72	21.23	0.002	66.47	64.16	0.020
	(0.33)	(0.32)		(10.19)	(9.97)		(22.57)	(17.88)	

Note: P-value from a difference in means test across genders. Earnings in thousands of dollars. SD reported in parentheses.

average, women believe that GPA at graduation in BEC and STEM are similar (p-value=0.497), while men believe that grades in STEM are lower than grades in BEC (p-value<0.01). As column (3) reflects, women believe that the GPA at graduation is higher than what men believe, regardless of major.

Regarding weekly study time, women's beliefs are 1-2 hours per week higher than men's. However, the pattern across majors is similar by gender. Both men and women believe that SSH is the major where students study the least per week followed closely by BEC, and STEM is the major that requires weekly study time (8-9 hours on average per week more than SSH).

In terms of earnings, participants believe that average earnings are higher in STEM, at around $64,000 - $66,000, followed by BEC at $54,000 - $55,000. SSH is in last place with average earnings beliefs around $40,000 -$41,000. As illustrated by the p-values in Table 2.2 column (9), women's beliefs about earnings are higher than men's regardless of major by about $1,500 - $2,000.

Differences between majors in beliefs about each attribute are further analyzed in Table 2.3. It presents the proportion of women and men that report each major having the highest attribute. For instance, 96% and 97% of men and women, respectively,

Table 2.3: Proportion of Participants that Rank a Major Highest for a Given Attribute, by Gender

	Av. GPA			Av. Study Time			Av. Earnings		
	Female	Male	P-value	Female	Male	P-value	Female	Male	P-value
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
SSH	0.33	0.46	0.005	0.02	0.02	0.995	0.01	0.01	0.443
BEC	0.26	0.22	0.397	0.01	0.03	0.058	0.12	0.11	0.235
STEM	0.41	0.32	0.043	0.97	0.96	0.180	0.87	0.89	0.174

Note: For each attribute and by gender, the table reports the proportion of participants that report each major having the highest level of the attribute (highest earnings, GPA or study time). For instance, 0.87 of women believe that earnings in STEM jobs are higher than in BEC and SSH, but only 0.01 believe that SSH jobs pay higher earnings than STEM and BEC. P-value from a difference in means test across genders.

believe that the average weekly study time is higher in STEM than in BEC and SSH. While only 2% of men and women believe that SSH majors require the highest study time. This suggests, that in general, students perceive STEM majors as requiring higher effort.

In terms of average earnings, 87% and 89% of women and men, respectively, believe that jobs in STEM areas pay on average higher earnings than SSH and BEC. However, 12-11% of women and men believe instead, that jobs in BEC pay higher wages than STEM and SSH. Only 1% of participants from each gender believe that SSH jobs pay higher earnings than the other two majors. These results imply that students expect big earning differences, particularly between SSH and the other areas.

Ranks in terms of average GPA at graduation are less extreme. A third of the women believe that average grades are higher in SSH, 41% believe that they are higher in STEM, and 26% believe that students in BEC graduate with the highest average grades. On the other hand, a higher share of men, 46%, rank SSH as having the highest grades at graduation, while 22% and 32% believe the same for BEC and STEM, respectively. Therefore, although there is a clear ranking of majors in terms of effort and earnings, the ranking is not as clear in terms of grades.

All the evidence in this section illustrates the variety of beliefs that students hold about major attributes, particularly in terms of grades across different majors. Beliefs about major characteristics like average grades, study time and earnings, tastes for each major, and shocks play a role when students decide to persist or switch out of a major. Since the administrative data do not provide information about any of these potential confounders, it has limited ability to shed light on the role that gender differences in grade sensitivity play in such decisions. Therefore, in the next section, I describe a survey experiment that allows me to quantify gender differences in grade sensitivity with a cleaner approach, by exogenously changing different major attributes, particularly average GPA at graduation.

2.5 Hypothetical Scenarios

To quantify gender differences in grade sensitivity, I use hypothetical scenarios to collect data that allows me to estimate students' preferences for different major attributes (Blass *et al.*, 2010; Delavande and Manski, 2015; Wiswall and Zafar, 2018; Folke and Rickne, 2022; Koşar *et al.*, 2022; Fuster and Zafar, 2023). Specifically, the survey included a hypothetical scenarios module that presented students with 10 different scenarios. In each scenario, majors were characterized by three attributes: average GPA at graduation, average weekly study time, and average earnings at a full-time job after graduation. Scenarios appeared one at a time. Table 2.4 is an example of how each scenario was presented to the participants.

I exogenously vary the magnitude of the attributes to identify participants' preferences for each of them. Scenarios are individual-specific to guarantee that each situation presented the student with attributes for each major that are realistic given the student's beliefs. Concretely, each scenario is a perturbation of the student's beliefs about the average GPA at graduation, study time, and full-time earnings for each major.

Table 2.4: Scenario Example

	Av. GPA	**Av. Study Hours per week**	**Av. Earnings after Grad. (full-time job)**
SSH	3.47	8.0	$24,000
BEC	2.23	7.0	$49,000
STEM	2.00	22.0	$46,000

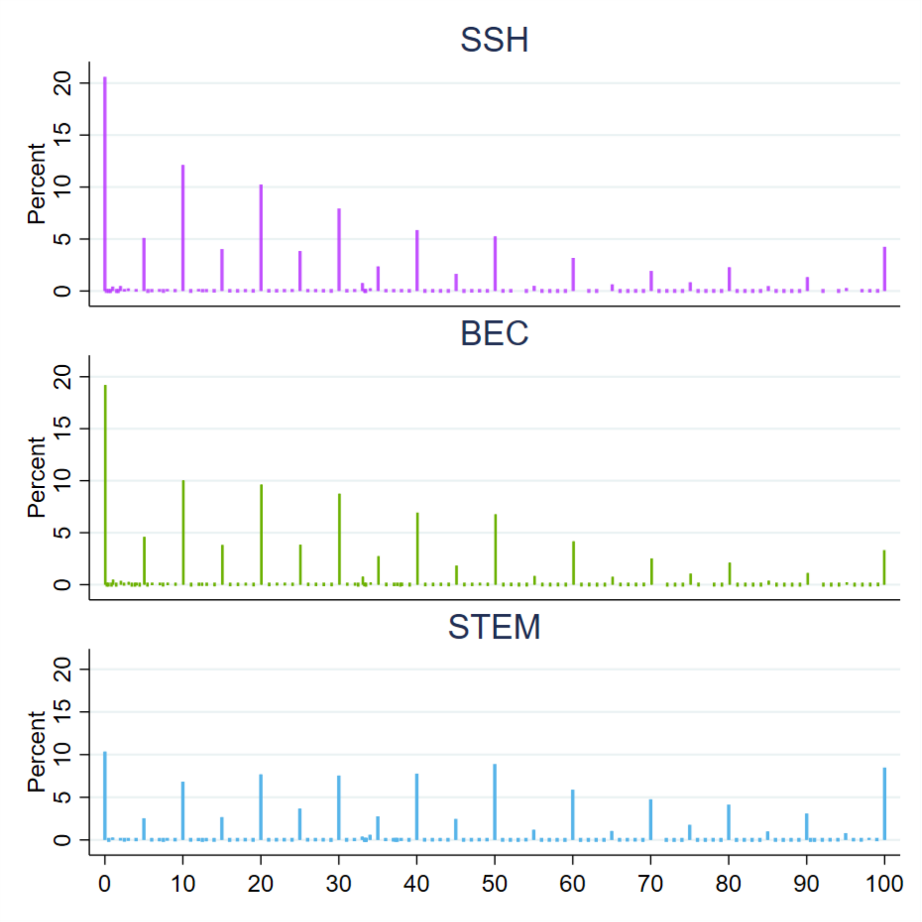
In each scenario, students reported the probability that they would choose each of the three majors given the characteristics.[9] Participants were asked to report probabilities because the scenarios they were facing were not fully specified. Majors can be characterized by more than the three attributes included in the survey. Therefore, participants are allowed to express their uncertainty about what they would choose given the incompleteness of the scenarios. Figure 2.2 shows the histogram of elicited choice probabilities for each major pooled across the ten hypothetical scenarios. As is common for probabilistic belief data (Manski, 2004), responses tend to be multiples of 5 and 10, which likely reflects minor rounding bias.[10] Figure 2.2 also shows that responses covered the whole support and not only values like 0, 50, or 100, which would reflect a problem with gross rounding (Manski, 2004). Additionally, 86% of the participants reported interior probabilities (not 0 or 100) in all their responses, which underscores the importance of allowing participants to express uncertainty in their choices.[11]

[9] The exact wording of the question was: *Imagine a situation in which you have not chosen a major yet and each major category is characterized as in the table below... What is the percent chance (or chances out of 100) that you would choose to graduate from each category given these characteristics?* See the survey instrument here for more details.

[10] Section 2.6 explains how the rounding bias is handled.

[11] Only 3% reported that they would choose one of the majors with 100% probability in all

Figure 2.2: Choice Probabilities by Major



Note: Histograms of choice probabilities for each major pooled across all scenarios.

An important implicit assumption when eliciting choice probabilities in this way is that stated choices reflect what the participants would choose in real-life scenarios. There is growing evidence that stated choices generated similar preference estimates as revealed preference approaches and that participants provide meaningful responses when the scenarios are realistic and relevant for them (Fuster *et al.*, 2021; Fuster and Zafar, 2023). In this case, major choice decisions are certainly relevant for college students. Moreover, as mentioned earlier, the scenarios were created to be as realistic

scenarios.

as possible given each participant's beliefs. Although I cannot test this assumption directly, it is reassuring that the results obtained from the hypothetical scenarios are consistent with the administrative data results in Figure 2.1 generated from actual choices, which suggests that participants provided meaningful responses (See section 2.6.2).

This design generates a panel of probability choices at the individual level, with 30 observations per participant, which allows me to estimate the distribution of preferences without any distributional assumptions. The next section describes the estimation procedure, and how the estimated preferences are used to calculate a measure of willingness-to-pay (WTP).

2.6 Preferences for Major Attributes

Similar to Wiswall and Zafar (2018), I use a simple model of expected utility of major choices that provides a framework to recover quantitative measures of WTP for the different major attributes. In particular, the model intends to recover how the utility of choosing a given major varies with GPA.

Let U_{ijs} denote the utility that student i gets from major j in scenario s. This utility is given by

$$U_{ijs} = X'_{ijs}\beta_i + \kappa_{ij} + \epsilon_{ijs} \tag{2.2}$$

where X_{ijs} is a vector that contains the attributes of the major: average GPA, average weekly study time, and the natural logarithm of the average earnings. κ_{ij} is a major-specific constant that captures tastes for the major.[12] Finally, since the scenarios in the survey are not fully specified, ϵ_{ijs} represents students' uncertainty about other attributes of the major at the time of the elicitation. I follow Blass *et al.* (2010) and

[12] For estimation purposes the constant for SSH major is normalized to zero, therefore the tastes for other majors are relative to SSH.

Wiswall and Zafar (2018) in interpreting ϵ_{ijs} as resolvable uncertainty, which means uncertainty at the time of the data collection that individuals know would be resolved in the case of an actual choice. The key identifying assumption is that, conditional on major, $\{\epsilon_{ijs}\}_{j=1}^{J}$ represents idiosyncratic variation which is orthogonal to the major attributes included in $\{X_{ijs}\}_{j=1}^{J}$.

Then, student i's reported probability of choosing major j in scenario s is

$$p_{ijs} = \int \mathbb{1}\left\{U_{ijs} > U_{ij's} \quad \forall j' \neq j\right\} dH_i(\epsilon_{is}) \tag{2.3}$$

where $H_i(\epsilon_{is})$ represents i's belief about the distribution of $\{\epsilon_{i1s}, ..., \epsilon_{iJs}\}$. I assume these beliefs are i.i.d Type I extreme value for all individuals. Therefore, student i's reported probability of choosing major j in scenario s takes the following form:

$$p_{ijs} = \frac{exp(X'_{ijs}\beta_i + \kappa_{ij})}{\sum_{j'=1}^{J} exp(X'_{ij's}\beta_i + \kappa_{ij'})} \tag{2.4}$$

Applying the log-odds transformation to equation (2.4) results in the linear model in (2.5).

$$ln\left(\frac{p_{ijs}}{p_{ij's}}\right) = (X_{ijs} - X_{ij's})'\beta_i + (\kappa_{ij} - \kappa_{ij'}) \tag{2.5}$$

As is common in the literature (Blass *et al.*, 2010; Wiswall and Zafar, 2018), I introduce measurement error to the model in (2.5) to account for the possibility of the minor rounding bias mentioned earlier. The assumption is that measurement error takes a linear-in-logs form, therefore the reported log-odds ratio is

$$ln\left(\frac{\tilde{p}_{ijs}}{\tilde{p}_{ij's}}\right) = (X_{ijs} - X_{ij's})'\beta_i + (\kappa_{ij} - \kappa_{ij'}) + \omega_{ijs} \tag{2.6}$$

where $\tilde{p}_{ijs}$ is the reported choice probability that measures the true probability, p_{ijs}, with measurement error ω_{ijs}. Additionally, the measurement error has a median of zero conditional on X.

Therefore, (2.6) is estimated using the Least Absolute Deviations (LAD) estimator. Since the left-hand side variable in (2.6) is the logarithm of the ratio of probability choices, extreme answers like 0 or 100 must be changed such that the natural logarithm is always defined. The LAD estimator has the advantage of not being sensitive to the values used to replace these extreme probabilities.[13] Variation in major attributes and variation in participant's choice probabilities across the 30 observations per respondent allows identifying the vector β_i for each student i separately. This allows for a non-parametric characterization of the preferences distribution.

2.6.1 Estimates of Preferences for Major Attributes

Table 2.5 reports the β_i estimates from equation (2.6), bootstrapped standard errors are reported in parentheses.[14] The first column shows the average estimate for each attribute and tastes across all individual-level estimates. Columns 2 and 3 report the average estimates by gender.

The average estimates have the expected signs: estimates for GPA at graduation and log of earnings are positive, while the estimates for study time are negative. This means that, on average, students prefer majors that pay higher earnings after graduation and have on average higher GPA, but lower weekly study time. By gender, the estimates for major attributes present the same qualitative patterns as the average estimates. Additionally, all attributes are statistically different from zero. In terms of tastes, on average, students prefer BEC and STEM majors less than SSH majors (the estimates are relative to SSH), although among men the average BEC and STEM

[13] Probabilities of 0 were replaced with 0.001 and 100 with 99.9.

[14] Sample size is smaller because seniors are not included in the analysis of the hypothetical scenarios data since they are closer to graduation and their preferences for major attributes might be different than those of less senior students. However, all results are qualitatively the same if seniors are included. Additionally, I drop outliers with WTP for study time or GPA greater (as defined in the next subsection) than \$100,000 or less than -\$100,000 (5.5% of the sample).

taste estimates are not statistically different from zero.

Table 2.5: Estimates of Preferences for Major Attributes

	Overall	Female	Male
	(1)	(2)	(3)
GPA at Grad.	0.650***	0.689***	0.574***
	(0.064)	(0.079)	(0.118)
Study time (h/week)	-0.070***	-0.060***	-0.090***
	(0.007)	(0.009)	(0.014)
Log earnings	4.569***	4.058***	5.558***
	(0.154)	(0.182)	(0.291)
Taste for BEC	-0.430***	-0.557***	-0.184
	(0.085)	(0.105)	(0.143)
Taste for STEM	-0.078	-0.244**	0.244
	(0.096)	(0.113)	(0.175)
N	1,192	786	406

Note: Table reports the average of the coefficientes across the relevant sample. Tastes for BEC and STEM are relative to SSH. Asterisks denote estimates that are statistically different from zero based on bootstrapped standard errors. *Significant at 10%, **5%, ***1%

Given the difficulty of interpreting the magnitudes in these estimates, the next sub-section converts the estimates to a willingness-to-pay (WTP) measure in order to quantify the gender gap in grade sensitivity in an easily interpretable way.

2.6.2 Willingness-To-Pay Measures

In this section, I calculate WTP measures based on the estimated preferences. These estimates translate the differences in utility due to different amounts of a given attribute into the earnings that would make the student indifferent between the two attribute levels.

The thought experiment to compute the WTP is as follows: consider a change in the level of attribute X_k from $X_k = x_k$ to $X_k = x_k + \Delta$ with $\Delta > 0$. Given the linear utility function, it is possible to write the following indifference condition in terms of earnings Y:

$$x_k\beta_{ik} + \beta_{i1}ln\left(Y\right) = \beta_{ik}\left(x_k + \Delta\right) + \beta_{i1}ln\left(Y + WTP_{ik}\left(\Delta\right)\right) \tag{2.7}$$

Solving (2.7) for WTP gives the following expression:

$$WTP_{ik}\left(\Delta\right) = \left[exp\left(\frac{-\beta_{ik}}{\beta_{i1}}\Delta\right) - 1\right] \times Y, \tag{2.8}$$

which is individual i's willingness to pay for a Δ increase in attribute k. Equation (2.8) depends on the ratio of the student preferences for attribute k, β_{ik}, and preferences for earnings, β_{i1}. Additionally, given the log form in the utility for earnings, the WTP measure depends on the level of earnings Y. For the calculations, Y is the average earnings across all participants across all scenarios ($53,318). The objective of having the same level for all respondents is that any gender differences in WTP discussed later will reflect only differences in preferences, not differences in earnings.

Table 2.6: WTP Estimates

	Dollars			% of Av. Earnigs			
	Overall	Female	Male	Overall	Female	Male	P-value[a]
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
GPA at Grad.	8,309	9,089	6,799	15.58	17.05	12.75	0.099
	[6,608]	[7,790]	[4,882]	[12.39]	[14.61]	[9.16]	0.018
	(652)	(811)	(1,126)	(1.22)	(1.52)	(2.11)	
Study time	-1,479	-1,428	-1,579	-2.77	-2.68	-2.96	0.725
	[-638]	[-608]	[-714]	[-1.20]	[-1.14]	[-1.34]	0.234
	(196)	(241)	(355)	(0.37)	(0.45)	(0.67)	
N	1,192	786	406				

Note: Table reports WTP mean, median in squared brackets, and bootstrapped standard errors in parentheses in dollars and as percentage of average earnings. All means and medians are statistically different from zero at 1%.
[a] P-value from a difference in means or medians test by gender.

Table 2.6 shows the average and median WTP measures for one extra unit of the attribute. That is one whole GPA point at graduation (from 2.3 to 3.3 for example) and one extra hour of study time per week. All means and medians reported in Table 2.6 are statistically different from zero (p-value<0.01). Columns (1)-(3) present the WTP measures in dollars and the last three columns display the WTP as a percentage of average earnings. The stars in the male columns (3) and (6) represent the significance level from a difference in means (or medians) test by gender.

On average, students are willing to pay 16% of the average annual earnings for a one-point increase in the average GPA at graduation of a given major but must be compensated with an extra 3% in average annual earnings to study one more hour per week. By gender, women are willing to pay 17% of their annual earnings for the one-point increase in the average GPA at graduation, but men only 13% (p-value< 0.1). However, there is no gender difference in the average WTP for weekly study time.

I interpret the WTP for GPA as a measure of students' sensitivity to grades. Since the objective is to understand why women and men value grades differently and how this could impact their major choices, I focus on this measure henceforth.

Table 2.7 reports the gender gap in WTP for GPA at graduation conditional on background characteristics. In particular, Table 2.7 reports α_1 from:

$$WTP_{GPAi} = \alpha_0 + \alpha_1 Female_i + \mathbf{C}_i + \xi_i \tag{2.9}$$

where the outcome variable is participant i's WTP measure for GPA at graduation. $Female_i$ is an indicator equal to one when the participant is female. $\mathbf{C}_i$ includes controls for family income, parents' education, minority status, SAT/ACT scores, school year, and indicators for honors students and majors.

Column (1) reports the overall conditional gender gap at \$3,057. This gap means that women are willing to forego \$3,057 of average annual earnings more than men for an extra GPA point at graduation in a given major. I interpret this difference

Table 2.7: Gender Gaps in WTP for GPA

	Overall	STEM/BEC	SSH
	(1)	(2)	(3)
Female	3,057**	3,760**	1,760
	(1,440)	(1,661)	(2,801)
Mean	8,309	9,414	6,307
R2	0.02	0.02	0.02
N	1,192	768	424

Note: Outcome variable is WTP for an extra point in av. GPA at graduation. All columns control for household income, parents education, SAT/ACT, school year, honors, minority. Additionally, column (1) controls for major. Bootstrapped standard errors reported in parentheses. Columns (2) and (3) split sample by reported major of participants. *Significant at 10%, **5%, ***1%.

as the gender gap in grade sensitivity since women are willing to "pay" more for the point increase. In columns (2) and (3) the sample is split by major: STEM/BEC versus SSH.[15] From this, it is clear that the overall gender gap is driven by the gap among STEM/BEC students where the difference in WTP for GPA at graduation between genders reaches $3,760. The gap is smaller ($1,760) and not statistically different from zero among the SSH students. These results are consistent with the administrative data evidence in Figure 2.1 discussed earlier: women in STEM/BEC majors are more sensitive to grades than men, but this gap is not observed in other majors.

2.7 What Could Be Driving the Gap?

There could be many potential mechanisms driving the gender differences in grade sensitivity documented in the previous sections. For example, the literature suggests gender differences in risk aversion (De Paola and Gioia, 2012), willingness to compete (Buser *et al.*, 2014), self-confidence (Ellis *et al.*, 2016; Moakler and Kim, 2014), and

[15] STEM and BEC majors are pooled together given the similar patterns in grade sensitivity observed in Figure 2.1.

beliefs about what it takes to graduate from a male-dominated major (Owen, 2020).

Another possibility is beliefs about gender discrimination and labor market standards (Steele *et al.*, 2002). There is evidence that women face different standards than men in hiring and promotion decisions, especially in male-dominated areas (Foschi *et al.*, 1994; Goldin and Rouse, 2000; Quintero, 2008; Williams *et al.*, 2014; Funk and Parker, 2018; Alam and Tapia, 2020). Thus, it is reasonable that female students could anticipate facing gender discrimination in the labor market, and even have heterogeneous beliefs about the level of discrimination they could experience in different fields. These beliefs could impact their response to grades and major choices, and help to explain the gender gap in sensitivity for grades documented earlier.

The primary focus of this section is on beliefs about gender discrimination and labor market standards for several reasons. First, although there is a substantial amount of research about gender discrimination in the labor market, there is considerably less work on anticipated discrimination and even less on its potential effects on major choices (Steele *et al.*, 2002; Alston, 2019). Therefore, studying this mechanism represents a significant contribution to our knowledge about the effects of gender discrimination in different spheres of life. Second, it is important to investigate mechanisms that do not rest on inherent differences between men and women (risk-aversion, self-confidence, willingness to compete), but instead rest on beliefs about the labor market, like anticipated gender discrimination, because providing evidence of their relevance would suggest different policy implications than other explanations. Third, I also collect data about self-confidence and beliefs about grades in different fields, however they do not seem to be systematically related to grade sensitivity. Therefore, I consider them later in section 2.8.

In the next section, I present a theoretical model of major choices that incorporates potential discrimination in the labor market against women to develop intuition about the role of beliefs about gender discrimination in decisions about field of study. Then, I provide evidence of the gender gaps in beliefs about gender discrimination and labor market standards using the survey data. Finally, I provide evidence of the importance of those beliefs in explaining the gender differences in grade sensitivity documented earlier.

2.7.1 Conceptual Framework

In this section, I setup a theoretical framework to formalize the intuition behind how beliefs about gender discrimination in the labor market can lead to gender differences in grade sensitivity and major choices. I add the employer side from Coate and Loury (1993)'s model of labor market discrimination to a framework where students revise their major decisions after receiving grades. When students decide to stay or leave a major they take into account their study costs, beliefs about their ability, and potential gender discrimination in the labor market.

I incorporate the possibility of gender discrimination by making the utility from each major depend of the probability of finding a job, which could differ by gender given how students believe the labor market works. They believe that as in Coate and Loury (1993), if employers in a given field discriminate against women they impose a more rigorous hiring rule for them. Female students incorporate that differential treatment in their major decision as a lower probability of getting hired in that field.

The goal is to show that by allowing major choices to be affected by discrimination in the labor market, women and men that receive the same grades make different decisions about staying in or leaving a given major. The difference arises because they believe they will be treated differently in the labor market based on gender.

Environment

Consider a mass one of female (F) students and a mass one of male (M) students. Gender is denoted by $g \in \{M, F\}$. There are two majors (k): STEM/BEC denoted by S and SSH denoted by N. All students are initially enrolled in major S. Students can be high (h) or low (l) ability, but they do not observe their level of ability. There is a P proportion of high-ability individuals. Additionally, students have a heterogeneous marginal cost for an extra hour of studying $c_i \sim U(0, 1)$.

Students receive grades which are noisy signals about their ability. Grades are drawn from $[0, 1]$ according to the pdf $f_h(\theta)$ if the student is high ability or $f_l(\theta)$ if they are low ability. The corresponding CDFs are F_h and F_l, respectively. I assume that $f_h(.)$ and $f_l(.)$ satisfy the Monotone Likelihood Ratio Property (MLRP).[16] Thus, higher grades are more likely if the student is high ability.[17]

Students believe there is a separate labor market for each field, which means that students who graduate with a degree in major k participate in the labor market for field k. Additionally, they believe that employers behave as follows. Employers in a given field have a prior belief π_g^k about the fraction of high ability individuals in the pool of workers of gender g. Employers get $x_h^k > 0$ if they hire a high ability student and $x_l^k < 0$ if they hire a low ability student. They observe students' GPA (grades) at graduation, θ, which are a noisy signal of the student ability, and update their beliefs about that particular student being high ability following Bayes rule. The posterior probability is denoted by:

[16] $\psi(\theta) = \frac{f_h(\theta)}{f_l(\theta)}$ is strictly increasing and continuous in θ for all $\theta \in [0, 1]$

[17] MLRP implies that F_h FOSD F_l, and that, for a given prior, the probability of being high ability is increasing in the grades (signal).

$$p(\theta;\pi_g^k)=\frac{\pi_g^k f_h(\theta)}{\pi_g^k f_h(\theta)+(1-\pi_g^k)f_l(\theta)} \tag{2.10}$$

Hiring Decisions

The firm will optimally choose to hire a student that provides signal θ if and only if

$$p(\theta;\pi_g^k)x_h^k-[1-p(\theta;\pi_g^k)]x_l^k \geq 0 \tag{2.11}$$

Using (2.10) in the condition above, a firm hires a student if and only if:

$$\frac{f_h(\theta)}{f_l(\theta)} \geq \frac{1-\pi_g^k}{\pi_g^k}\frac{x_l^k}{x_h^k} \tag{2.12}$$

The MLRP implies the existence of a unique $\tilde{\theta}_g^k \in (0,1)$ such that (2.12) holds with equality.[18] This means that the employer follows a cutoff hiring rule. The firm will hire a student if their grade (signal) is higher than the cutoff, i.e. $\theta > \tilde{\theta}_g^k$.

Assume that $\frac{x_l^N}{x_h^N} < \frac{x_l^S}{x_h^S}$, i.e. the ratio of profit to losses is higher in S than in N. This is reasonable since the potential problems of hiring a low-ability worker in a more technological sector like S might be greater than in N. This assumption guarantees that the signal cutoff in the N sector is lower than in the S sector, $\tilde{\theta}_g^N < \tilde{\theta}_g^S$, for both genders. This conclusion is consistent with the average response in the survey about the labor market standards.

Additionally, it is the case that $\frac{d\tilde{\theta}_g^k}{d\pi_g^k} < 0$.[19] As the belief about the proportion of high ability workers in the pool of potential employees increases the firm uses a lower threshold for the grades in order to hire them; in other words a less rigorous standard. This property will be relevant later when considering perceived discrimination in the

[18] If (2.12) does not hold with equality for any $\theta \in (0,1)$, then $\tilde{\theta}(\pi_g^k)=0$ if $\frac{f_h(0)}{f_l(0)} = \frac{1-\pi_g^k}{\pi_g^k}\frac{x_l^k}{x_h^k}$ or $\tilde{\theta}(\pi_g^k)=1$ if $\frac{f_h(1)}{f_l(1)} = \frac{1-\pi_g^k}{\pi_g^k}\frac{x_l^k}{x_h^k}$. See Fang and Moro (2011) for more details.

[19] See Fang and Moro (2011) for proof.

S labor market. Discrimination will be introduced as the belief that the proportion of high-ability (productive) women in the S labor market is lower than men, $\pi_F^S < \pi_M^S$. Therefore, $\tilde{\theta}_F^S > \tilde{\theta}_M^S$, which means that in the presence of perceived gender discrimination women will face a more rigorous standard than men since they need to provide a better signal (higher GPA) in order to get hired.

Revising Major Decisions

At the end of their first year, students receive their grades, θ_i, drawn from their respective distribution according to ability, $f_h(.)$ or $f_l(.)$. Given this new information, students update their beliefs about being high-ability following Bayes rule, and potentially revise their major choice. P is the proportion of high-ability students and the prior belief about being high-ability.

Given grades, θ_i, the posterior belief about being high-ability is

$$P'(\theta_i) = \frac{Pf_h(\theta_i)}{Pf_h(\theta_i) + (1-P)f_l(\theta_i)} \tag{2.13}$$

After receiving the grades and updating their beliefs, students compare the utility of each major and choose the one with higher utility. Therefore, based on the new information they can stay in S or switch into N.

The utility of studying each major is given by the expected payoff of a job after graduation in that field minus the cost of studying. Jobs in S pay 1 and jobs in N pay $v < 1$. The expected payoff depends on the probability of finding a job, which depends on the probability that the GPA at graduation, θ_i, is above the cutoff in the corresponding field, $\tilde{\theta}_g^k$. This probability is:

$$P'(\theta_i)[1 - F_h(\tilde{\theta}_g^k)] + (1 - P'(\theta_i))[1 - F_l(\tilde{\theta}_g^k)] \tag{2.14}$$

Utilities for each major are as follows:

$$U_g^N(\theta_i) = vP'(\theta_i)[1 - F_h(\tilde{\theta}_g^N)] + v(1 - P'(\theta_i))[1 - F_l(\tilde{\theta}_g^N)] - \delta^N c_i \tag{2.15}$$

$$U_g^S(\theta_i) = P'(\theta_i)[1 - F_h(\tilde{\theta}_g^S)] + (1 - P'(\theta_i))[1 - F_l(\tilde{\theta}_g^S)] - \delta^S c_i \tag{2.16}$$

where δ^k represents the number of study hours required by major k. Major S requires more study time than N, $\delta^S > \delta^N$.

A student i of gender g chooses to stay in S if $U_g^S(\theta_i) \geq U_g^N(\theta_i)$. The MLRP implies that a reservation grade $\theta_i^* \in (0, 1)$ exists such that[20]

$$\begin{cases} U_g^S(\theta_i) \geq U_g^N(\theta_i), & \text{if } \theta_i \geq \theta_i^* \\ U_g^S(\theta_i) < U_g^N(\theta_i), & \text{otherwise} \end{cases} \tag{2.17}$$

Thus, a student decides to leave S if their grade is not high enough relative to their reservation grade θ_i^*.

It is the case that

$$\frac{\partial \theta_i^*}{\partial \tilde{\theta}_g^S} > 0 \tag{2.18}$$

which means that the higher the cutoff grade to get a job in S the higher the reservation grade to stay in S. The reservation grade is a function of both labor market cutoffs, $\tilde{\theta}_g^N$ and $\tilde{\theta}_g^S$, the payoff v in field N, the grade θ_i, the cost of studying c_i, and the study time in both majors δ_S and δ_N.

Anticipated Gender Discrimination in S

Consider the case in which female students expect to face gender discrimination in the labor market for major S. That means that they assume that in the S labor market employers believe that there is a higher proportion of high-ability men than women, $\pi_F^S < \pi_M^S$. Given that employers follow a cutoff hiring rule (See 2.7.1), women believe they will face a higher cutoff than men in S labor market in order to get a full-time job, i.e. $\tilde{\theta}_F^S > \tilde{\theta}_M^S$.

[20] See Appendix C for proof.

Then, for an identical man and woman (same c_i and ability), and given (2.18)

$$\theta_i^*(\tilde{\theta}_F^S) > \theta_i^*(\tilde{\theta}_M^S) \tag{2.19}$$

This means that the woman requires a higher grade than the man to stay in S. In other words, if they both receive the same grade θ_i, such that $\theta^*(\tilde{\theta}_F^S) > \theta_i > \theta^*(\tilde{\theta}_M^S)$, then the man is going to stay in S and the women is going to leave S (switch to N). Notice that this is consistent with the patterns in grade responsiveness from Figure 2.1, where at every grade level women are more likely than men to switch out of STEM/BEC majors. Additionally, this framework provides a compelling explanation for how anticipated discrimination can affect students' WTP for grades differently depending on gender as discussed in section 2.6.2.

2.7.2 Anticipated Gender Discrimination: Empirical Evidence

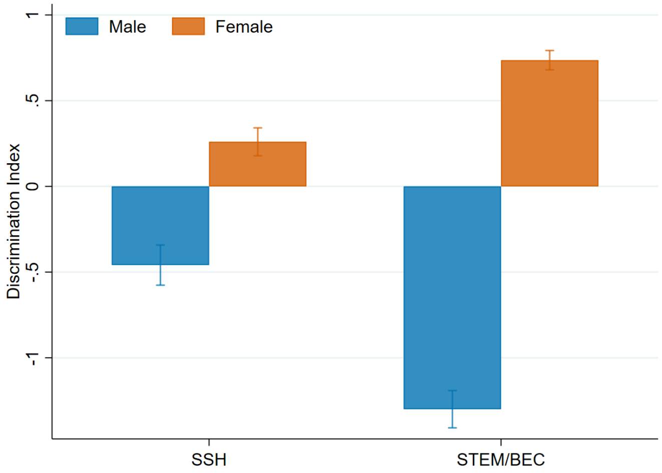
In this section, I document gender differences in students' beliefs about gender discrimination and hiring standards in the labor market using the survey data, and present evidence of the importance of those beliefs to understand the gender differences in grade sensitivity.

In order to measure beliefs about anticipated gender discrimination in the labor market, participants responded to a gender discrimination panel in the survey. They were asked, how likely (on a 5-point Likert scale) it would be that: (1) it is harder to find a job because of their gender, (2) their supervisor/boss would treat them differently because of their gender, and (3) their peers/coworkers would treat them differently because of their gender.[21] Given that beliefs about discrimination can be different for different majors or fields, the questions were asked for each major separately. Their responses for each major were combined using Principal Components

[21] Given the leading nature of these questions they were asked at the end of the survey.

Analysis (PCA) to create a major-specific index of anticipated gender discrimination.[22]

Figure 2.3: Gender Discrimination Index by Gender



Note: Average gender discrimination index for each major by gender. The index calculated using PCA and the responses to how likely (on a 5-point Likert scale) it would be that: (1) it is harder to find a job because of their gender, (2) their supervisor/boss would treat them differently because of their gender, and (3) their peers/coworkers would treat them differently because of their gender. Spikes represent 95% CI.

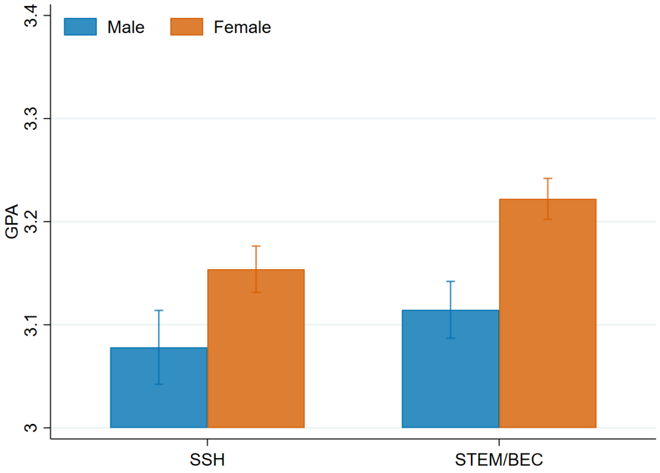
Figure 2.3 shows the average gender discrimination index by major and gender. By construction, each index has a mean of zero (and standard deviation of one), therefore negative (positive) numbers imply anticipated gender discrimination that is lower (higher) than average. Men anticipate facing less discrimination due to their gender in both fields than the average participant. The story is different for the female students. Female participants foresee facing more gender discrimination than average in both fields. However, women anticipate that they will face more gender discrimination in the STEM/BEC labor market than in the SSH labor market (p-value<0.01). This result is consistent with evidence of higher difficulties in the labor

[22] All the results are qualitatively consistent if the major-specific indexes are constructed with a PCA algorithm that takes into account the discreteness of the variables.

market for women in male-dominated fields (Foschi *et al.*, 1994; Goldin and Rouse, 2000; Funk and Parker, 2018; Alam and Tapia, 2020).

As the theoretical framework shows, a way in which discrimination could affect women's decisions is through beliefs that they need to provide more or better evidence of competence than men in order to be hired, especially in male-dominated fields. Therefore, participants were asked to report what they think is the minimum GPA at graduation that they will require to secure a full-time job in STEM/BEC (SSH) if they were to graduate with a degree in STEM/BEC (SSH). Each participant answered the question for each major, regardless of the major they report to be enrolled in.

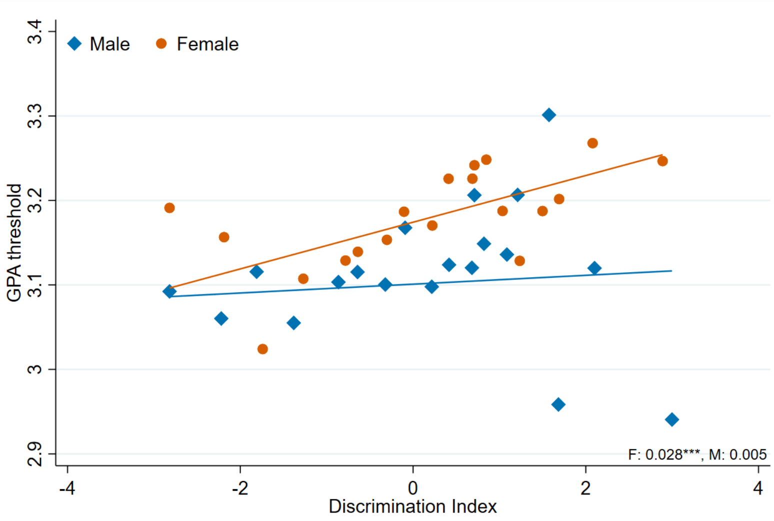
Figure 2.4: Average Beliefs about Min. GPA Necessary for Full-Time Job in Given Field



Note: Average belief about the minimum cumulative GPA at graduation required to secure a full-time job in each field by gender. Spikes represent 95% CI.

Figure 2.4 shows the average GPA threshold for each major by gender. In general, participants believe they would need a lower GPA to secure a job in SSH than in STEM/BEC. For instance, on average women believe they would need a GPA 0.068 higher to get a job in STEM/BEC than in SSH (p-value<0.01). On average men believe they would need 0.036 extra GPA points at graduation to secure a job in STEM/BEC instead of SSH (p-value=0.014). Moreover, on average, women believe they need a higher GPA than men to secure a full-time job, regardless of the major they graduate from. The gender gaps in beliefs about the GPA necessary to get an SSH or STEM job are 0.075 and 0.11, respectively (p-value<0.01 for both).[23] In summary, women believe they will need to provide a better signal of their competence in the labor market in the form of a higher GPA than their male counterparts, especially in order to secure a job in the STEM/BEC field.

Figure 2.5: Discrimination and Thresholds Relationship



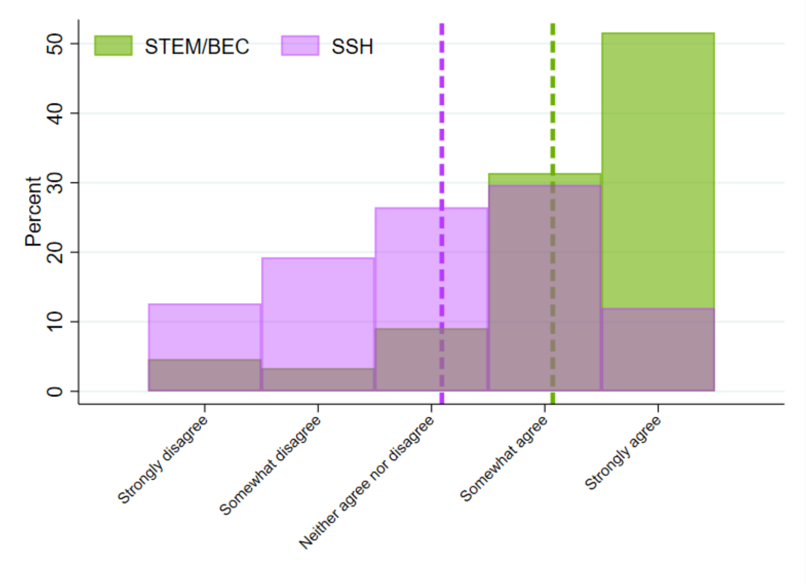
Note: Markers are from a binned scatter plot between GPA thresholds to get a full-time job and the anticipated discrimination index. Lines are fitted values from a regression of the GPA threshold on the discrimination index separately by gender and standard errors are clustered at individual level. Coefficients at the bottom left corner are the slopes of each line. *Significant at 10%, **5%, ***1%.

The binned scatter plot in Figure 2.5 shows the relationship between beliefs about anticipated gender discrimination in the labor market and beliefs about the GPA required to secure a job. There is a positive and significant relationship (p-value< 0.01) between the level of discrimination that a woman believes she is going to face and

[23] These gender gaps are not statistically different from each other (p-value=0.25).

her beliefs about the minimum GPA at graduation required to secure a full-time job. However, this positive relationship is weaker for men, which is not surprising since men expect to experience less gender discrimination. Therefore, their beliefs about the GPA required to get a full-time job are not as strongly related to discrimination as they are for women.[24]

Figure 2.6: Female Participants Agreement with "Women Need a Higher GPA to Compete Against Similar Man", by Major



Note: For each major, histogram of female participants responses to "How much you agree with: A woman competing for a job in this field would need a higher GPA than an otherwise similar man to be competitive." Dashed lines represent the average level of agreement by major.

The fields where women expect to face more discrimination due to their gender are the fields in which they foresee they will need to provide a really strong signal about their ability in order to be competitive. In fact, I asked female participants how much they agree (on a 5-point scale) with the idea that a woman applying for

[24] However, there is a statistically significant positive relationship for men when dropping eight outlier observations with high discrimination index but low labor market standards.

a job after graduation in a given field would need a higher GPA than an otherwise similar man to be competitive. Figure 2.6 summarizes the responses. The dashed lines represent the average response per major. On average, the level of agreement with the idea of women requiring a higher GPA in order to be competitive is higher if applying for a STEM/BEC job than an SSH job (p-value<0.01). Moreover, almost 83% of the female participants somewhat agree or strongly agree with the statement in the case of a STEM/BEC job, whereas only 42% agree to the same extent in the case of an SSH job. These results reinforce the previous conclusions, women believe they will have a harder time in the STEM/BEC labor market.

Table 2.8 analyzes the role of anticipated discrimination and beliefs about GPA thresholds to secure a full-time job in explaining the gender gap in grade sensitivity. The first column duplicates column (1) in Table 2.7, which reports the conditional average gender gap in WTP for GPA at graduation, $3,057. Column (2) controls for beliefs about the necessary GPA to get a full-time job in STEM/BEC and SSH fields. Although the gender gap is still statistically different from zero, the point estimate decreases by 13% ($2,671). Therefore, beliefs about facing different standards in the labor market seem important to understand why women and men value grades differently.

Discrimination also plays a role in explaining the gender gap in WTP aside from its effects through the GPA thresholds as can be seen in column (3), which controls directly for the anticipated discrimination indexes in STEM/BEC and SSH. In this case, the gender gap is no longer statistically different from zero and the point estimate decreases by 36% to $1,965. This reduction suggests that anticipated discrimination is relevant for understanding the gender gap in grade sensitivity. Finally, column (4) includes controls for both discrimination indexes and GPA thresholds. In this case, the point estimated decreases to $1,600, a 48% reduction, and it is not statistically significant.

Table 2.8: Importance of Anticipated Discrimination and GPA Thresholds for the Gender Gaps in WTP for GPA

	(1)	(2)	(3)	(4)
Female	3,057**	2,671*	1,965	1,600
	(1,440)	(1,431)	(2,085)	(2,045)
Belief GPA Threshold STEM/BEC		5,548**		5,533**
		(2,433)		(2,406)
Belief GPA Threshold SSH		-499		-487
		(2,195)		(2,057)
Anticipated Discrimination STEM/BEC			620	613
			(729)	(742)
Anticipated Discrimination SSH			-261	-270
			(535)	(541)
Mean	8,309	8,309	8,309	8,309
R2	0.018	0.024	0.019	0.025
N	1,192	1,192	1,192	1,192

Note: Outcome variable is WTP for an extra point in av. GPA at graduation. All columns control for household income, parents education, SAT/ACT, school year, honors, minority, and major. Bootstrapped standard errors reported in parentheses. *Significant at 10%, **5%, ***1%.

These results suggest that beliefs about anticipated gender discrimination and labor market standards are important to understand why women and men value grades differently. Additionally, they support the intuition formalized in the conceptual framework that highlights the role of beliefs about anticipated discrimination in women's major choices.

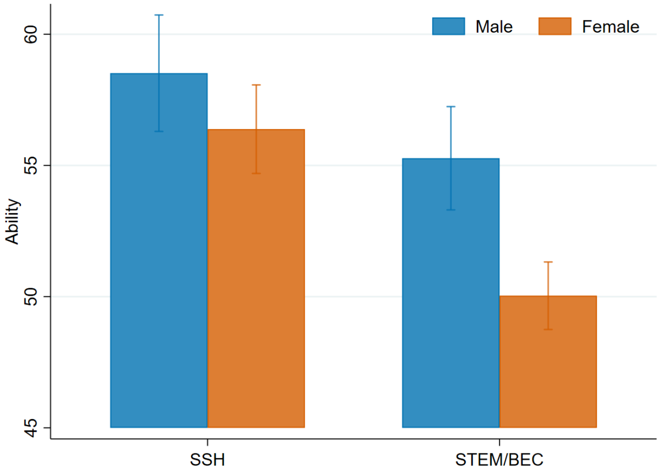
2.8 Other Explanations

Aside from beliefs about gender discrimination and labor market standards, there could be other mechanisms that contribute to explaining the gender gap in sensitivity to grades. The literature suggests that gender differences in self-confidence, and beliefs about the grade distribution in different fields could play a role in this context. In this section, I discuss the empirical evidence of their contribution to explaining why women and men react differently to grades using the survey data.

2.8.1 Self-Confidence

There is evidence that women are less confident in their quantitative abilities than men. For example, Ellis *et al.* (2016) finds that women that take Calculus I start and end the term with less confidence in their mathematical abilities than men. Similarly, Moakler and Kim (2014) finds that women report lower academic and mathematics confidence than men, and this is related to their lower chances of choosing a STEM major. Therefore, women could interpret less-than-stellar grades in STEM and BEC majors as confirmation of their lack of ability and subsequently switch out of them.

Figure 2.7: Average Beliefs about Ability in Each Major



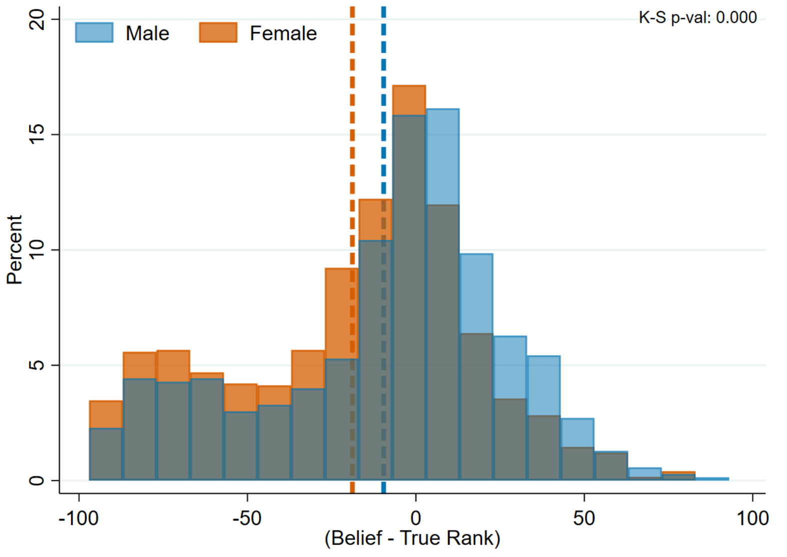
Note: Average ability ranking in each major by gender. Rank is on a 1-100 scale where higher numbers represent higher ability. Spikes represent 95% CI.

In the survey, participants report beliefs about their SSH and STEM/BEC ability as their rank relative to peers on a 1-100 scale.[25] Figure 2.7 reports the average rank by gender and major. Students report higher beliefs about their ability in SSH than in STEM/BEC: on average women (men) report rankings 6.35 (3.24) points higher in

[25] The higher the number the better the ability relative to peers.

SSH than in STEM/BEC (p-value<0.01 for both genders). On average, men report higher beliefs about both their SSH and STEM/BEC ability than women. However, only the gender gap in beliefs about STEM/BEC ability is statistically different from zero (p-value <0.01).

Figure 2.8: Ability Over/Under Confidence, by Majors



Note: Histogram, by gender, of the difference between participants' beliefs about their rank in their reported major and their "true" rank in that major based on reported cumulative GPA. Dashed lines represent the mean of each respective distribution. K-S p-val: p-value from a Kolmogorov-Smirnov test for the equality of the distributions.

Figure 2.8 plots the distribution of the difference between participants' beliefs about their rank in their reported major and their "true" rank in that major (Belief - True Rank). True rank is calculated using the administrative data of students registered in each of the majors during the Spring of 2021. Specifically, in the administrative data, all students in a major cohort are ranked based on their cumulative GPA and this ranking is used to assign the true rank to the survey participants based on the cumulative GPA they provided. Then, this difference (Belief - True Rank) is the error in participants' beliefs about their ability. If the error is positive (negative) participants are over (under) confident in their ability.

In Figure 2.8, the vertical dashed lines represent the mean of the distribution by gender and show that on average participants are under-confident in their ability. In other words, participants report a worse rank than their actual position based on their GPA. However, women are more under-confident than men as illustrated by the lower mean (p-value <0.01), and the extra mass below zero in the female histogram.[26,27]

Table 2.9: Importance of the Errors in Beliefs about Ability for the Gender Gaps in WTP for GPA

	(1)	(2)
Female	3,057**	2,905**
	(1,440)	(1,439)
Error in Beliefs about Ability		-18
		(20)
Mean	8,309	8,309
R2	0.018	0.019
N	1,192	1,192

Note: Outcome variable is WTP for an extra point in av. GPA at graduation. All columns control for household income, parents education, SAT/ACT, school year, honors, minority, and major. Bootstrapped standard errors reported in parentheses. *Significant at 10%, **5%, ***1%.

Table 2.9 examines the role of over/under confidence in the gender differences in grade sensitivity. Column (1) reproduces the first column from Table 2.7, which reports the conditional average gender gap in WTP for GPA at graduation, $3,057. Column (2) controls for the error in beliefs about ability as described before: belief - true rank. This error reduces the gender gap slightly (5%), but it remains statistically significant.

Despite the fact that the gender differences in self confidence have the expected

[26] Based on the Kolmogorov-Smirnov test, female and male distributions are statistically different from each other in both panels of Figure 2.8 (p-value<0.01).

[27] Results are qualitatively the same if the distributions are analyzed separately for students enrolled in SSH and STEM/BEC. See Figure B.1 in the Appendix.

patterns, theses results do not support the role of self-confidence as an important driver of the gender differences in sensitivity to grades.

2.8.2 Beliefs about Grade Distribution in Different Fields

Academic performance is one of the main reasons for changing majors (Wright, 2018). However, there is evidence that students sometimes hold erroneous beliefs about the grade distributions in different fields. For example, Owen (2020) finds that men are more likely to underestimate the median grade of students enrolled in STEM majors, while women overestimate it. If women overestimate the grades of the students graduating from STEM or BEC majors, they might believe that their less-than-stellar grades in the introductory classes are not good enough to succeed in those majors, and they might switch out. Therefore, erroneous beliefs about the grades at graduation in different majors seem like a potential explanation for the gender gap in grade sensitivity.

Table 2.10: Average Beliefs about GPA at Graduation, and Actual Average GPA at Graduation by Gender and Major

	Actual	Beliefs		
	GPA	Female	Male	p-value
	(1)	(2)	(3)	(4)
SSH	3.38	3.46	3.36	0.000
STEM/BEC	3.43	3.37	3.23	0.000

Note: Column (4) is the p-value of a difference in means test across genders within major, columns (2) and (3).

In the survey participants are asked to report what they believe is the average GPA of students who graduate from each major. Table 2.10 reports the average response for each major by gender in columns (2) and (3). The fourth column reports the p-value from a difference in means test between genders. Regardless of major, women believe that the average GPA at graduation is higher than what men believe. All participants believe that the average GPA at graduation is lower among STEM/BEC students.

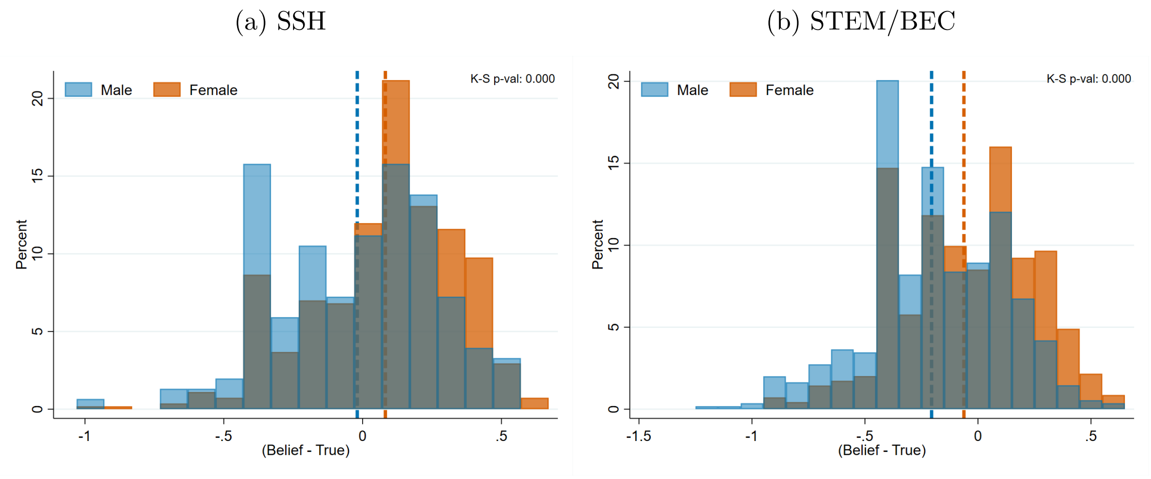
It is important to learn how close these beliefs are to the actual GPA of people graduating from each major. To do so, I use the administrative data described in section 2.2. In column (1) Table 2.10 reports the average GPA among the students that graduate from each major during the Spring of 2019.[28] Opposite to what participants believe, the average GPA of students graduating with a STEM/BEC degree is slightly higher than the GPA of students graduating with a SSH degree. Additionally, the average male belief in STEM/BEC is statistically lower than the actual GPA of people graduating with that degree (p-value<0.01). Although the average male belief about the grades in SSH is slightly lower than the actual average GPA in this major, the difference is not statistically significant (p-value=0.230). On the other hand, the average female belief for SSH is statistically higher than the actual GPA for SSH (p-value<0.01), and the opposite is true for STEM/BEC (p-value<0.01).

To further analyze the error in these beliefs, Figure 2.9 shows the distribution of the difference (error) between a participant's belief about the GPA at graduation for the major they report to be enrolled in and the corresponding average GPA at graduation from the administrative data (Spring 2019). Negative (positive) numbers indicate that participants underestimate (overestimate) the GPA at graduation. The dashed lines represent the mean of each distribution.

On average, men in SSH majors tend to underestimate the grades of their graduating peers. However, on average, women in SSH hold correct beliefs about the GPA

[28] I use Spring 2019 instead of Spring 2020 or Spring 2021 because those are semesters affected by different grading policies implemented as a response to the COVID-19 pandemic. However, results are qualitatively the same if any of those semesters is used instead.

Figure 2.9: Error in Beliefs about Av. GPA at Graduation, by Majors


(a) SSH

(b) STEM/BEC

Note: Histogram, by gender, of the difference (error) between a participant's belief about the GPA at graduation for the major they report to be enrolled in and the corresponding average GPA at graduation from the administrative data (Spring 2019). Dashed lines represent the mean of each respective distribution. K-S p-val: p-value from a Kolmogorov-Smirnov test for the equality of the distributions.

of their graduating peers.[29] In STEM/BEC, both women and men underestimate the GPA of the students graduating from those majors. Nonetheless, on average, women underestimate the grades of their graduating peers less than men (p-value<0.01 from a one-sided test). This means that women's beliefs are closer to the actual GPA of the Spring 2019 graduating class than men's.

Additionally, the distributions in each panel of Figure 2.9 are statistically different across genders as the p-value from the Kolmogorov-Smirnov test shows. Regardless of major, a higher share of women tends to overestimate the GPA of their graduating peers, as illustrated by the extra mass above zero in the female distributions relative to men's.

Table 2.11 evaluates the role that over or underestimation of the GPA at graduation plays in explaining the gender gap in grade sensitivity. Column (1) reproduces

[29] The mean for women is statistically not different from zero, p-value=0.4540

Table 2.11: Importance of the Errors in Beliefs about GPA at Graduation for the Gender Gaps in WTP for GPA

	(1)	(2)
Female	3,057**	2,796*
	(1,440)	(1,446)
Error in Beliefs about GPA at Graduation		1,871
		(2,529)
Mean	8,309	8,309
R2	0.018	0.019
N	1,192	1,192

Note: Outcome variable is WTP for an extra point in av. GPA at graduation. All columns control for household income, parents education, SAT/ACT, school year, honors, minority, and major. Bootstrapped standard errors reported in parentheses. *Significant at 10%, **5%, ***1%.

the first column in Table 2.7, which reports the conditional average gender gap in WTP for GPA, $3,057. Column (2) controls for the errors in beliefs about the GPA at graduation (the errors plotted in Figure 2.9). In this case, the gender gap estimate decreases by about 9% to $2,796.

These results do not provide strong support in favor of the hypothesis that holding erroneous beliefs about what is required to graduate from a given major is an important driver of the gender differences in grade sensitivity.

2.8.3 Relative Importance of Anticipated Gender Discrimination

The previous results imply that self-confidence and erroneous beliefs about grades in different fields do not play a significant role in explaining the gender gap in WTP for GPA. However, that analysis is done for each hypothesis independently. In this section, I provide suggestive evidence of the importance of anticipated gender discrimination relative to the other two explanations.

Table 2.12 analyzes the effect of each potential mechanism in explaining the gender gap in grade sensitivity when considered together. Column (1) presents the overall conditional average gender gap in WTP for GPA, $3,057. Column (2) controls for the error in beliefs about ability as described in section 2.8.1. Therefore, it reproduces column (2) in Table 2.9, where the point estimate decreases by about 5%. In column (3), I control for the errors in beliefs about the GPA at graduation in a given major as described in section 2.8.2. This reduces the point estimate by about 8%. This decrease is similar, in percentage terms, to the decrease in the gender gap in WTP for GPA when this hypothesis is considered separately. Lastly, column (4) controls for beliefs about anticipated gender discrimination. These measures decrease the estimated gender gap in WTP for GPA by around 46% relative to column (3), and the coefficient is no longer statistically different from zero. Overall, the gender gap in WTP for GPA decreases by 47%, from $3,057 to $1,437 when including all the controls.[30]

Table 2.12: Relationship between Gender Gaps in WTP for GPA and Pontential Mechanisms

	(1)	(2)	(3)	(4)
Female	3,057**	2,905**	2,664*	1,437
	(1,440)	(1,439)	(1,528)	(2,029)
Error in Beliefs about Own Ability		✓	✓	✓
Error in Beliefs about GPA at Graduation			✓	✓
Anticipanted Discrimination				✓
Mean	8,309	8,309	8,309	8,309
N	1,192	1,192	1,192	1,192

Note: Outcome variable is WTP for an extra point in av. GPA at graduation. All columns control for household income, parents education, SAT/ACT, school year, honors, minority, and major. Bootstrapped standard errors reported in parentheses. *Significant at 10%, **5%, ***1%.

The results in table 2.12 indicate that anticipated gender discrimination is, quantitatively, playing a more important role than the other two mechanisms given that its inclusion generates a greater decrease in the estimated gender gap in WTP for GPA. The results from an Oaxaca-Blinder decomposition further support this conclu-

[30] See Table B.1 in the Appendix for the coefficients of each of the different mechanisms.

sion.[31] 52% of the observed gender gap in WTP for GPA can be explained by gender differences in beliefs about anticipated gender discrimination, 7% is due to gender differences in self-confidence, and another 6% is attributed to gender differences in erroneous beliefs about the GPA at graduation from different majors.

2.9 Conclusion

The probability of women continuing their studies in or switching out of male-dominated fields like STEM and Economics depends more on their performance in relevant courses at the beginning of their college career relative to men. This paper studies why women and men react differently to grades during college and how this behavior impacts their decision to persist or switch out of a given major. Understanding why talented women with the potential to succeed in male-dominated fields drop out because of less-than-stellar grades in an introductory class is important for closing the gender gap in these areas, improving the labor market outcomes of highly skilled women, and achieving an efficient allocation of resources across fields of study and occupations.

Using administrative data from Arizona State University, I document gender differences in reaction to grades among undergraduate students. I find that among STEM and business students, the gender gap in the probability of persisting in those majors is negatively related to first year GPA.

The limited ability of the administrative data to shed light on the reasons that lead to those patterns, I use novel data from an online survey to quantify students' sensitivity to grades, and investigate the reasons why women and men react differ-

[31] The Oaxaca-Blinder decomposition allows me to determine the portion of the gender difference in WTP for GPA that is explained by group differences in the level of observable explanatory variables, in this case for each of the three mechanisms considered (Rahimi and Nazari, 2021).

ently to grades. I estimate students' grade sensitivity using the hypothetical scenarios methodology. I find that, conditional on background characteristics, women are willing to pay about $3,000 more of average annual earnings than men for a one-point increase in the average GPA at graduation in a given major. This gender gap is primarily concentrated among STEM and business students.

I provide evidence that anticipated discrimination in the labor market of male-dominated fields is important to understand this gender gap in grade sensitivity. I find that women believe that they are more likely to experience gender discrimination in the labor market than men, particularly in STEM and business fields. Additionally, I find that women believe that they will face a higher standard in the labor market in terms of GPA in order to get a full-time job. I provide evidence that the beliefs about higher standards are related to beliefs about gender discrimination in the labor market. Furthermore, my results show that beliefs about gender discrimination in the labor market account for 48% of the gender gap in sensitivity to grades.

Also, I propose a theoretical framework that formalizes the intuition about how these beliefs can lead to the gender differences in persistence observed in STEM and business majors. I show that by allowing major choices to be affected by discrimination in the labor market, women and men that receive the same grades make different decisions about staying in or leaving a given major. The difference arises because they believe they will be treated differently in the labor market based on gender.

I acknowledge that there are other mechanisms that could contribute to explaining the gender differences in grade sensitivity that I document. However, anticipated discrimination represents an explanation not often considered, and my results provide evidence of its importance in this context. In fact, considering the role of such beliefs is crucial to designing policies that effectively encourage the participation of women in traditionally male-dominated fields. For example, if students' beliefs about gender discrimination are close to the reality of the labor market, then policymakers should aim to solve the discrimination issues in the labor market. On the other hand, if students hold inaccurate beliefs about the labor market, information interventions could be a valuable tool. Therefore, assessing the accuracy of these beliefs represent an important avenue for future research.

Chapter 3

A (DYNAMIC) INVESTIGATION OF STEREOTYPES, BELIEF-UPDATING, AND BEHAVIOR (WITH KATHERINE COFFMAN AND BASIT ZAFAR)

3.1 Introduction

Many decisions, especially those regarding investments in human capital or workplace, are dynamic in nature. Take the case of a student deciding on what field to specialize in. After having taken courses in different fields, she receives noisy feedback about performance in them. She might update her beliefs in the moment, but then has time to process feedback and subsequently determines her future course of action. The fact that individuals usually receive feedback in real time but then have time to process it, and are rarely required to make a decision at the moment, is pervasive across many facets of daily life. For example, a worker, before applying for a promotion or a different job opportunity, has ample time to process the feedback that she has received up to that point in time. While there is a large literature that investigates immediate response to feedback, we know less about how responses to feedback may evolve *over time.*

This is the question that we set out to investigate in this chapter, with a focus on gender gaps. Ex-ante, there is reason to believe that gender differences in response to feedback may differ, since women are significantly less likely to opt into competitive tournaments than men (Niederle and Vesterlund, 2007), and also tend to be less over-confident than men on average (Barber and Odean, 2001). The response to feedback is also likely to differ by domain, since we know these gender gaps are more likely to manifest themselves in more male-typed domains (Beyer, 1990; Lundeberg *et al.*, 1994; Beyer and Bowden, 1997; Beyer, 1998; Coffman, 2014; Exley and Kessler, 2022).[1] It is, however, less clear how gender gaps in response to feedback may evolve *over time*.

The interest in this question is not merely an academic exercise. There is growing evidence that information interventions can be successful in debiasing individuals' beliefs and, in some cases, shifting their choices (see Roth *et al.*, 2021, and references, therein and Benjamin, 2019, for a review of belief updating in response to feedback in the laboratory). However, the potential of this path for reducing gender differences depends upon how men and women respond to feedback, specifically about their own abilities and talents. In particular, if it is the case that there are gender differences in how individuals respond to feedback in the moment, or in what kind of feedback is recalled and incorporated into beliefs and choices in the longer run, this could limit the effectiveness of information in closing gender gaps in educational and career choices.

We explore this set of open questions in a controlled, laboratory-style online experiment which is *dynamic* in nature. Importantly, the set-up allows us to generate exogenous variation in feedback to explore how individuals update their beliefs and

[1] Subsequently, a growing strand of empirical work has identified these differences in competitive preferences and overconfidence as factors in gender gaps in educational and career outcomes. For example, Buser *et al.* (2014) find that willingness to compete explains a significant portion of secondary school students' choices about whether to pursue the more demanding, and lucrative, math and science educational tracks. Reuben *et al.* (2017) find that competitiveness and overconfidence predicts earnings' expectations among college students. Reuben *et al.* (2019) find that competitive preferences can explain about 10 percent of the gender gap earnings at the time of college graduation. They find that overconfidence is also related with earnings, but the relationship varies over the life-cycle. More recently, Cortés *et al.* (2021) find that gender differences in overconfidence have gendered implications for the job search behavior of college students.

choices in response to good or bad news, over time.

Our experiment consists of two sessions, one week apart. In the first session, participants take two incentivized assessment quizzes in Round 1, one in math and one in verbal skills. Next, the participant reports her (incentivized) beliefs about absolute and relative Round 1 performance in each domain. We next inform participants that they will take a second round of quizzes one week later, and that these quizzes will be harder (mimicking the fact that tasks become more complicated in the real world as one progresses). Then, we elicit a series of choices about how they would like to be compensated for this future performance, knowing that one of these choices will be implemented to determine their Round 2 compensation. First, they choose between being paid for math performance under a piece-rate scheme or for verbal performance under a piece-rate scheme ($1 per correct answer). Next, we elicit their willingness-to-accept (WTA) competition in each domain using price lists. Participants make a series of choices between receiving either $1 per correct answer in verbal (math) or entering a competitive pay scheme in math (verbal). The competitive option pays $X per correct answer in math (verbal) *if* they place in the top 40% of performers in the Round 2 math (verbal) quiz, but 0 otherwise. For each domain, we vary X from $1.5 to $4 across the rows. We refer to these as the "Initial" decisions – beliefs and choices prior to receiving any feedback.

Treated participants – who constitute 82% of our sample – then receive feedback about their relative performance; the remaining 18% are the control which allows us to control for other time-varying factors unrelated to the feedback. For each of the domains, the computer randomly selects an individual from a peer reference group, and the participant learns if they performed better or worse than that individual. In this way, feedback is informative – a Bayesian should update their beliefs in response to whether they performed better or worse than a randomly drawn peer – but also noisy. In addition, *conditional on performance*, whether someone receives good news (that is, they performed better than an individual in the reference group) or bad news is random. This random variation is key for identification; it also provides a realistic degree of ambiguity for our participants. In our setting, there is still ample scope for self-serving interpretations of feedback, e.g. “Maybe I got unlucky in the peer that was drawn."

For half of our sample that receives feedback, we elicit beliefs and choices again immediately after the receipt of this feedback. These are what we refer to as the “Immediate” beliefs or decisions. The other half of our sample that receives feedback leaves the first session without providing updated beliefs or choices. All participants return for the second session one week later. In the second session, we again elicit the same beliefs and choice measures (the “Week After” decisions) from all participants, including from the control respondents who do not receive any feedback in the first session. All participants then take the two Round 2 quizzes. We also ask treated participants to recall the feedback they received in each domain at the end of the second session.

As mentioned above, our experiment is inspired by many settings, where individuals receive noisy feedback in different domains/tasks, and then decide what to specialize or compete in. The key feature that distinguishes our setup from most existing work on feedback is that we see how the response to feedback evolves over time (beyond the immediate impact that is typically observed in one-session experiments). We use a stylized, controlled environment to mimic important features of this setting, producing several advantages. First, we observe individual measures of ability in both domains. Second, we observe exogenous changes in the individual’s information set (which are quite hard to isolate in non-experimental settings), allowing us to cleanly study belief updating. Third, we have precise measures of beliefs. And, finally, we have well-defined measures of payoffs for the chosen domain as well as for the counterfactual domain - this offers us an advantage since counterfactual payoffs are, by definition, not observed in the field. Our design, by necessity, is a stripped down version of real settings: for example, treated individuals receive only one signal in each domain, and the signal is about relative ability. Enriching the design – by providing additional signals and/or signals about absolute ability – would require a much larger sample size.

Our design allows us to collect detailed information about beliefs, choices, and recall at different points in time in both a female and a male-typed domain. This allows us to ask whether there are differences across men and women and/or differences across the associated stereotype of the task. Thus, we present results in terms of two gender gaps: the male $-$ female gap (average differences between men and women) and the gender-congruence gap (average differences between individuals in the gender-congruent domain and individuals in the gender-incongruent domain).[2] Both gaps are potentially important for understanding gender disparities in educational and career settings of interest.

Over 1,800 Arizona State University undergraduates participated in our experiment. In line with past work, we find significant gender gaps in beliefs and choices at baseline. On average, men are more overconfident than women, with a larger male-female gap in math than in verbal. These beliefs are highly predictive of choices about how to be compensated in the second round, even controlling for measured ability. Men are significantly more willing than women to choose to compete in math, but not in verbal.

We find that feedback has a sizable, significant impact on individuals' beliefs and

[2] Concretely, this congruence gap compares the decisions of men in math together with the decisions of women in verbal to decisions of men in verbal and women in math.

choices. Immediately after receiving feedback, individuals revise their beliefs and choices by between $0.15 - 0.35$ standard deviations (SDs) on average. But, by one week later, these revisions partially fade back toward starting points. The impact of bad news seems to fade *less* over time than the impact of good news, particularly for women (relative to men) and for individuals who receive bad news in incongruent domains (relative to congruent domains). We can also compare men and women's reactions to the Bayesian benchmark. Consistent with the literature, we see that individuals under-react to feedback on average. In addition, the under-reaction is observed for men and women for both kinds of news, good and bad. Thus, both men and women are under-responding to bad news, relative to the Bayesian benchmark. Men are simply under-responding more.

Before feedback, gender gaps conditional on measured performance (both the male $-$ female gap and the gender-congruence gap) are significant. Women's beliefs are approximately 0.35 SDs more pessimistic than men's, and women are approximately 0.15 SDs less willing to compete. We also document significant differences by gender congruence. Individuals are 0.15 SDs more confident in congruent domains and are 0.20 SDs more willing to compete. Immediately after feedback, gender gaps are somewhat reduced, particularly for beliefs. However, in the week following feedback, gaps grow back toward their starting point. In particular, gender gaps in choices one week later are indistinguishable from gender gaps at baseline.

We show that the persistence of gender gaps in our setting is driven largely by reactions to bad news. Conditional on having the same performance, having made the same initial decisions, and receiving *positive* feedback, there are no gender gaps in beliefs or choices immediately after or one week after feedback. If anything, women have more optimistic beliefs than men one week later. Put differently, men and women seem to respond similarly to positive feedback conditional on having the same starting point. Individuals also update beliefs and choices similarly in response to positive feedback across congruent and incongruent domains.

On the other hand, gender does seem to play a role in how individuals update their beliefs and choices in response to *negative* feedback. If we take a man and a woman with the same performance and the same initial beliefs, then provide the same bad news, the woman holds more pessimistic beliefs about herself one week later compared to the man. Similarly, even if we hold fixed performance and initial choices, women (compared to men) are less willing to compete one week after bad news.

We also document differences in how choices respond to bad news across congruent and congruent domains. If we compare two individuals with the same performance and who made the same initial choices, the individual who received bad news in the incongruent domain is less willing to compete one week later than the individual who received bad news in the congruent domain.

These results are not driven by forgetting of feedback. Overall, 88 percent of feedback is accurately recalled one week later. We find that women are significantly more likely to accurately recall feedback than men. Both men and women are significantly more likely to remember bad news than good. But these differences do not explain the persistence of gender gaps that we observe; we estimate similar results among the subset of individuals who accurately recall their feedback.

Our results have several implications. While we show that individual beliefs and choices can be meaningfully shifted by provision of information, the impact of feedback on gaps is more limited. Furthermore, the impact of feedback seems to at least partially fade out over time, with beliefs and behavior moving back in the direction of initial decisions. This suggests that a better understanding of how initial beliefs and choices are formed, absent feedback, is crucial to uncovering the sources of these "sticky" gender gaps.

Our dynamic setting allows us to highlight that, even over a short window of time, the impact of feedback can change.[3] In particular, we see evidence that the impact of bad news fades over the course of week for men, but not women, and in congruent domains, but not incongruent domains. As a result, gaps one week later are larger than gaps immediately after feedback. Our finding that there are significant gender gaps in decisions after bad news, even conditional on performance, feedback received, and initial decisions, point toward the challenge of addressing gender gaps through information interventions. Differential reactions to the same information can exacerbate initial gaps.

3.2 Related Literature

A growing body of research uses controlled experiments to better understand how beliefs respond to feedback (see Benjamin, 2019 for overview of belief updating literature), with a few offering insights on gender differences (Mobius *et al.*, 2021, Ertac and Szentes, 2011, Coutts, 2019, and Shastry *et al.*, 2020). There seems to be evidence that women may update their beliefs more conservatively, particularly in more male-typed domains. There is also evidence on how information can shift competitive preferences; Cason *et al.* (2010), Ertac and Szentes (2011), Wozniak *et al.* (2014), and Shastry *et al.* (2020) highlight that providing feedback about performance can

[3] It could be that in the week between the two sessions, following the feedback received in the first session, treated respondents' interest in a given subject may endogenously evolve. For example, following positive feedback in a domain, respondents may go out and seek more information about that domain or their interest in it may increase. While we do not observe what exactly transpires during the week, our effects will include the impact of any of these subsequent endogenous behavioral responses.

reduce gender gaps in competitive tournament entry in laboratory settings.[4] Closest to our work is Coffman *et al.* (2021), who study how men and women update their beliefs in response to feedback on absolute ability, comparing reactions across male and female-typed domains. We build on this prior work by linking beliefs to choices and by exploring differences in recall of feedback and updating over time.

One focus of this literature has been investigating asymmetries in belief updating. While some have found evidence of motivated updating (greater adjustment to good news than bad - see, for instance, Eil and Rao, 2011; Mobius *et al.*, 2021; Charness and Dave, 2017; and the dynamic setting of Zimmermann, 2020), others have not (Ertac and Szentes, 2011; Grossman and Owens, 2012; Schwardmann and Weele, 2019; Gotthard-Real, 2017; Barron, 2021; Coutts, 2019; and Coffman *et al.*, 2021). These types of motivated responses to feedback have been a focus of psychology work on this topic, with many studies documenting that individuals more often attribute positive feedback to internal factors (i.e., their own talent), and negative feedback to external factors such as bad luck (Heider, 1958, Miller and Ross, 1975, Campbell and Sedikides, 1999, and Mezulis *et al.*, 2004).

This emerging literature largely studies immediate responses to feedback; information is received and beliefs are updating over the course of a single experimental session. But, for many real-world applications, significant time may pass between the moment feedback is received and the time at which important choices are made. Consider the question of what kind of education to pursue– an individual may have a prior belief about her abilities, and then receive noisy feedback about her true talents over time. Choices about which field to major in, or what kind of career to enter, likely occur weeks, months, or years after the provision of feedback. This makes it

[4] Kessel *et al.* (2021) find that information on the gender gap in willingness to compete can also reduce the gap in tournament entry.

essential to consider not only how beliefs and choices respond to feedback immediately after its provision, but also to understand how the impact of feedback changes over time. Could it be the case that certain kinds of feedback are more likely to be recalled, or more likely to have a lasting impact on choices? And what role do those types of differences have in contributing to the persistence of gender gaps?

Mobius *et al.* (2021) invite participants back to the laboratory one month after their feedback intervention and elicit decisions about whether to compete. They find that beliefs – shaped by the exogenous feedback intervention – strongly predict decision-making, suggesting a persistence of the impacts of feedback over time. Zimmermann (2020) further unpacked the dynamics of belief updating in a setting with feedback. Specifically, in his setting, participants take an IQ test and receive feedback on their performance. He finds that, one month later, beliefs are more responsive to positive than negative feedback, and positive feedback is more likely to be accurately recalled, consistent with theoretical models of motivated reasoning. This complements evidence from other settings in which people seem to have overly positive memories of past events. In a controlled experiment, Chew *et al.* (2020) find that, months after taking an IQ test, participants have self-serving beliefs about their own performance on specific questions. In the field, Huffman *et al.* (2019) find that, even in the face of high incentives, managers have over-confident beliefs about past performance and consistently over-predict future performance. Recently, economists have made significant advances in incorporating models of memory and associative recall to explain the formation and persistence of biased beliefs (Bordalo *et al.*, 2020, and Enke *et al.*, 2020).

These results suggest that introducing a dynamic element may have significant implications for how feedback is processed and incorporated. Over time, there may be a larger role for biases, including gender biases, in shaping beliefs and choices.

Understanding exactly how these dynamic features interact with gender is one of the important open questions that we address.

In the domain of education, some field experiments have investigated how relative performance feedback affects beliefs and/or choices (for example, see Azmat and Iriberri, 2010; Franco, 2019; Bobba and Frisancho, 2021; Owen, 2020). Some of these papers also document effects that vary by gender. Relative to this literature, our major innovations are that we elicit beliefs *and* choices, at multiple points in time, with a focus on gender differences. We investigate how the effects of information dissipate over time, and how the fade-out (if any) may depend on the stereotype of the domain and gender.

Finally, our finding of gender differences in reactions to bad news is consistent with a growing body of work exploring the persistence of men and women after bad news, losses, or failures. Pairing a laboratory experiment with evidence from a large-scale math competition in the field, Buser and Yuan (2019b) show that women are less likely than men to choose to compete again after a loss. Subsequent work has found similar results across a range of field settings of interest, including college entry exams, conference submissions, math competitions, and politics (Kang *et al.*, 2021, Fang *et al.*, 2021, Pereda *et al.*, 2020, Brown *et al.*, 2019, Wasserman, 2020, and Ellison and Swanson, 2018). Relatedly, Gill and Prowse (2014) find that women exert less effort after competitive losses relative to men, and Shastry *et al.* (2020) find that women are more likely to attribute negative feedback to ability rather than bad luck.

Our work highlights that reactions to feedback depend not only on gender, but also on the domain: stereotypes matter in predicting how individuals update their beliefs and choices in response to feedback. In addition, our setting allows us to explain *why* such gendered patterns may emerge: our results suggest that differential belief updating about oneself, particularly over time and especially in response to negative feedback, may play a critical role in driving these gender gaps. It is worth noting that we do not find that either gender is over-reacting to information (relative to a Bayesian benchmark). If anything, both genders seem to under-react.

3.3 Study Design and Administration

3.3.1 Experimental Design

We study the evolution of choices and beliefs over time by conducting two online sessions, one week apart. Each session consists of a performance component – solving verbal and math quizzes – as well as elicitation of beliefs and choices. Participants are told that, at the end of Session 2, one of the two rounds of performance quizzes will be selected at random for payment. Figure 3.1 shows an overview diagram of the experimental design described in this section.

Session 1

Round 1 Performance Quizzes: In Session 1, participants start by taking two Round 1 performance quizzes: a math and a verbal quiz. The order in which the two quizzes appear is randomized across subject. Each quiz consists of 12 multiple choice questions, ordered randomly, with one question appearing at a time. Participants are allowed a maximum of 30 seconds to attempt each question, reducing the chances that they look up answers on the internet. The quizzes include modified questions from the GRE, SAT, and a logic book (Russel and Carter, 2001). If Round 1 is randomly chosen for payment, participants receive $1 per correct answer in one of the two quizzes, chosen at random.

We study both math and verbal because we aim to understand the role of domain stereotypes in driving beliefs and choices. Participants perform in two domains with different gender stereotypes, the more male-typed math domain, and the more female-typed verbal domain. Indeed, 72% of our sample say that women have an advantage in the verbal domain, 70% say that men have an advantage in the math domain; the majority – 58% of our respondents – believe that both these statements are true.[5] But, we take care to assure similarity across the domains in other dimensions, including average difficulty, question style, and reference group. This allows us to better isolate the associated stereotype of the domain, something that is difficult to do in the field.

When designing the quizzes, we tested a large battery of verbal and math questions on Amazon Mechanical Turk (MTurk). Then, informed by this pilot data, we constructed two sets of Round 1 quizzes, one harder and one easier. Within each difficulty level, the quizzes were designed such that we expected average absolute performance to be similar across math and verbal. In this way, we reduce the chances that observed differences across domain are due to differences in difficulty of the quizzes, rather than differences in the domain. By choosing two levels of difficulty, we can also ask directly whether, within domain, the exogenously-assigned level of difficulty is relevant for beliefs and choices.

Participants are randomized into a difficulty level for both quizzes at the beginning of the experiment (with an equal chance of being assigned to either level), and were not aware of this feature. Because our main results do not depend on the randomly

[5] We should note that, despite these perceptions, men actually outperform women in the quizzes in both domains in our experiment (see our results section below). However, in our view, these perceived gender advantages, consistent with previous work (Bordalo *et al.*, 2019), suggest that we indeed achieved at least some across-domain variation in perceived gender-type. It is after all the *perceived* gender stereotype, more so than actual differences in performance, that matter for understanding the impact of stereotypes on choices and beliefs.

assigned difficulty level (that is, there are no significant interactions between gender, gender congruence of the domain, and the exogenously assigned difficulty level), we simply pool the two difficulty levels together for our main analysis and include an easy/hard indicator in our specifications.

Initial Beliefs about Round 1 Performance Quizzes: Following their completion of the Round 1 quizzes, participants report their beliefs about their Round 1 performance in both domains, math and verbal. Note that participants complete all beliefs questions for one domain, then all beliefs questions for the second domain. For each domain, there are four beliefs questions. First, we ask participants to guess their absolute score – their total number of correct answers on the quiz. Incentive compatibility is ensured by offering $1 if their guess is correct. We also ask them about how confident they are in their guessed score: that is, what are the chances that you earned exactly that score? We apply the incentive-compatible belief elicitation procedure used by Mobius *et al.* (2021), implemented as in Coffman *et al.* (2021). As an example, for these two questions, a participant might tell us they believe they had a score of “8," and that they think there is a 75% chance that they had exactly a score of “8."

Next, participants provide beliefs of relative performance in Round 1: specifically, participants are asked to consider how their performance on each quiz compares to the performance of a reference group. This is a group of 9 individuals from the same population that took the same quiz as the participants but prior to the full roll-out of the experiment.[6] First, we ask them what they believe their rank position is, 1-10, when compared to the reference group, 1 being the best position. We incentivize participants by offering them $1 if their guess is correct. Second, to obtain a full

[6] There is a reference group for each difficulty level in Round 1. Therefore, participants are compared to the reference group matching their randomly assigned difficulty level in Round 1.

prior belief distribution, for each possible position (1-10) in the ranking, we also elicit participants' beliefs about the likelihood that they ranked in each position when comparing their performance to the reference group. We again use the incentive procedure of Mobius *et al.* (2021). For the analysis, we invert the rankings such that a higher rank means a better rank, with 10 being the best rank.

We elicit an extensive set of beliefs, covering both absolute and relative performance. This is helpful in understanding choices, for which beliefs of both absolute and relative performance are relevant. We also elicit full subjective belief distributions because it allows us to construct Bayesian benchmarks for belief updating. We should note that we elicit beliefs about Round 1 performance, and elicit choices for preferred compensation in Round 2. In this way, participants are asked to use feedback on past performance to make decisions for the future, mimicking a feature of many contexts of interest.

Initial Choices for Round 2 Compensation: After the beliefs elicitation section, we inform participants that they will take a second round of quizzes a week later, during Session 2, and that the quizzes will be harder on average than in Round 1. While participants have to take both quizzes, they have a choice of how they want to be compensated for their performance (that is, if Round 2 is randomly chosen for payment). We ask them to make a series of choices between pairs of payoff schemes. One of the options always involves being paid for verbal performance, and one of the options always involves being paid for math performance. We vary the particulars of the payment schemes across choices.

First, we ask participants to choose between piece-rates: would they rather be paid piece-rate for their Round 2 math performance or piece-rate for their Round 2 verbal performance (each $1 per correct answer)? Then, we use two price-lists, one for each domain, to elicit their choices over competitive payment schemes. Figure 3.2

shows the price list for math. The "first option" offers $X per correct answer in the math quiz if the participant performs in the top-4 when compared to the reference group in terms of Round 2 math performance, 0 otherwise. We vary X from $4 to $1.5 as one proceeds down the six rows on the price list. The "second option" always offers $1 per correct answer on the verbal quiz. We are essentially asking, how much does the reward for successfully competing in math have to be to induce a participant to choose math over a piece-rate verbal scheme? The price list for verbal is analogous (see Figure D.1), with the first option offering $1 per correct answer in the math quiz and the second option offering different rewards along the six rows that range from $1.5 to $4 if the participant performs in the top-4 in the verbal quiz, 0 otherwise. Participants know that one of all the decisions made during the experiment about how they want to be compensated for Round 2 will be randomly chosen to calculate their earnings if Round 2 is chosen for payment. We included two understanding check questions in each of the price lists to ensure that participants understood the payment mechanisms.

From the price lists, we calculate a willingness to accept (WTA) competition in each of the domains for each participant. This is the lowest dollar amount (X) at which the participant prefers the competitive payment scheme in that domain to the piece-rate scheme in the other domain. If the participant always chooses the competitive scheme, we set the WTA to $1.5. On the other hand, if they always pick the fixed reward of $1, we set the WTA to $4.5. For participants with multiple switch points in the price list, we code the WTA as missing. In the main text, we focus on WTA as our choices outcome. In Appendix E, we present corresponding results for choosing math in the choice between the two piece-rate schemes.[7]

[7] The main conclusions of the chapter, that feedback has a limited impact on closing the gender gap in the choice of math holds for this analysis as well.

The beliefs and choices reported at this stage are referred to as "Initial" in the analysis.

Feedback Provision: After making the choices about Round 2, a subset of respondents – specifically 82% – receive feedback. The remaining 18% form the Control group (C). Participants in the feedback groups receive a noisy signal about their relative performance. For each of the domains, the computer randomly selects an individual from the reference group, and the participant learns if they performed better or worse than that individual. Ties are broken randomly.[8]

We provide only one signal per domain in our study, simplifying the implementation and analysis. While it is always challenging to extrapolate, we think it is likely that our results from a single signal setting are likely to provide valuable insights into reactions to information in contexts where more than one signal is available. Understanding how multiple simultaneous or sequential signals interact with gender differences is an important topic for future work.

Immediate Beliefs and Choices: Within the feedback group, half of the participants (those assigned to the Immediate group) answer the exact same belief elicitation questions again immediately, within Session 1. They are also asked again about how they would like to be compensated for their Round 2 performance, answering the exact same questions again. We refer to these beliefs and choices that are elicited immediately after feedback as "Immediate" in the analysis. Participants in the control group, as well as those who receive feedback but are not randomly assigned to the Immediate group, do not see these questions a second time within Session 1.

In designing the experiment, we were unsure whether being required to provide beliefs and choices immediately after feedback would "anchor" participants to a certain set of posterior beliefs one week later. For this reason, we randomly assign only some

[8] The feedback order for the two domains is randomized.

of our treated participants to the group where there is an elicitation immediately after feedback, allowing us to provide an empirical answer to this question. As it turns out, there are no significant differences across the Immediate and non-Immediate feedback groups in beliefs and choices one week later, which suggests that having to incorporate the feedback immediately through a belief and choice elicitation in Session 1 does not change the dynamic impact of feedback in our setting.

Session 1 concludes for all participants with a survey section where we ask them some demographic questions: gender, race, household income, parents educational attainment, high school GPA, high school rank, college GPA, major, school year and a survey measure of risk aversion.

Session 2

Session 2 occurs one week later. Seven days after the completion of Session 1 participants received an email with the access link for Session 2.[9] We do not have insight into what students may do, look up, or think about in the week between feedback provision and Session 2. To the extent that differences in behavior in the interim contribute to our results, we think these forces are likely be relevant in other contexts as well.

Week After Beliefs and Choices: Session 2 starts with all participants, including the control group, the Immediate group, and the non-Immediate group, answering the belief elicitation questions a final time and making their choices about how they

[9] They were told the link would remain active for 24 hours. A first reminder was then sent the next day to the participants who had not completed Session 2 during the allotted time. The reminder gave an extra 24 hours to complete Session 2. A final reminder was sent the morning of the following day to participants. Thus, participants were effectively given 72 hours to open the link and complete Session 2.

prefer to be compensated if Round 2 is chosen for payment. Again, these questions are identical to what they have seen previously. These are referred to as the "Week After" beliefs/choices in the analysis.

Round 2 Performance Quizzes: Next, participants complete the Round 2 math and verbal quizzes. The format is exactly as in Round 1, except that the questions in the second round are, on average, harder than those in Round 1, as participants are told to expect. Note that independent of assigned Round 1 difficulty, all participants take the same harder quizzes in Round 2.

Conclusion of Session 2 and Assessment of Recall: At the end of the session, we ask participants their perceptions of the gender stereotype of each domain by asking them to assess which gender they think knows more about each of the domains on average: men or women. This concludes the experiment for the Control group. Additionally, participants in the feedback group are asked to recall the feedback they received a week before in each domain. They receive $0.25 for each piece of feedback correctly recalled.

Importantly, the control of an experiment allows us to shutdown some problematic selection effects. We observe men and women in both domains, across both rounds of performance. This allows us to compute key counterfactuals, including their counterfactual earnings under different choices about compensation schemes. In addition, we observe the feedback that the individual receives, and we can take advantage of exogenous variation (since, conditional on performance, whether someone gets a positive or negative signal is random); changes in information sets are difficult to fully observe in the real world. Even when such changes are observed, they tend to be endogenous which limits the inference from such variation.

Our experiment was created using oTree (Chen *et al.*, 2016). Online Appendix A shows the screenshot of the experimental instructions. We registered the experiment during the data collection for Session 1, prior to looking at any data (AEA RCT Registry "A Dynamic Investigation of Stereotypes, Belief Updating, and Behavior", ID AEARCTR-0005712; web link: https://www.socialscienceregistry.org/trials/5712). We registered the design, plan for determining sample size, and primary outcomes of interest, but we did not pre-specify a specific analysis plan.[10]

3.3.2 Sample

The experiment was run at Arizona State University (ASU), one of the largest public universities in the United States, during April 2020. It was advertised as a two-session experiment, scheduled one week apart. We guaranteed a completion payment of $12.50 with the possibility of additional incentive pay ranging between $0 and $49.5. The guaranteed payment as well as any additional compensation were only paid out after the completion of the second session, with hopes of minimizing attrition. Students were directly invited to participate via email. We initially targeted students in the Honors College at ASU – a selective, residential college that recruits academically outstanding undergraduates across the nation – via a weekly email digest sent out by the college. We then also advertised the study on the MyASU website, accessible only through a student's ASU ID and password, broadening our reach to all ASU students.

In order to reach a target of 1,800 completes for Session 2 (as mentioned in our plan, registered at the start of Session 1), we targeted roughly 2,000 completes for

[10] We did note that we planned to use the control group to check for time trends and that we would not focus on the control group for our main analysis. Because we unexpectedly did find some time trends, the control group ends up serving as another important reference group in our study. In the analysis below, we are explicit about when the control group is included or not, and what the reference group is for all comparisons.

Session 1. A total of 2046 students completed Session 1 and 1816 completed Session 2 a week after. Our analysis sample consists of these 1816 participants. Our attrition rate is low, which we believe is partly a result of back-loading the compensation. Importantly, it does not differ by gender, performance in Session 1, the treatment group the participant is assigned to, initial beliefs, or the feedback that one receives (see Table D.1). We do, however, find that Honors College students were less likely to attrit.

Women make up 60% of our sample. Although women are over-represented in our sample relative to ASU's student population (49% female), there is no differential selection on observables across genders (see Table D.2). Panel A of Table 3.1 reports the gender-specific means of different characteristics of our sample, and the third column reports the p-value of a difference in means test. Relative to men, women in our sample are more likely to be Hispanic and first generation students, and have lower average family incomes, but similar gendered patterns are observed in the underlying population (Table D.2). In line with existing evidence, we also see that women report a significantly higher level of risk aversion than men (3.66 versus 3.31, on a 1-7 Likert-scale).

The average (median) time taken to complete Session 1 was 40.5 (28) minutes. The corresponding statistics for Session 2 were 37 (17). There is no gender difference in the average time taken to complete either session. The average (median) earnings for men were $19.8 ($19), and for women were $19.4 ($19); the p-value for a test of the equality of the average earnings across gender is 0.024.

3.4 Results

We present the results in several parts. We start by documenting the beliefs and choices at the Initial stage. We then show how they evolve over the course of the experiment, comparing initial beliefs and choices to those that are elicited immediately after feedback and those that are elicited one week later. We then consider whether these patterns vary according to gender, the gender stereotype of the domain, or the type of news received. We close by considering the implications of these results for the persistence of gender gaps.

3.4.1 Descriptive Statistics

We first summarize the baseline data in Panel B of Table 3.1. On average, men perform better than women in both domains in both rounds. As expected, average performance levels are substantially lower in the second round.[11] The full histogram of Round 1 ranks by domain and gender are presented in Appendix Figure D.2. For individuals who rank first or last in a domain, we cannot interpret their feedback as randomly-assigned. These individuals are included in our analysis presented below. But, we note that our main results, presented in Table 3.4, hold when we exclude these individuals at extreme ranks (see Appendix Table D.3).

Initial Beliefs and Choices

Both men and women, on average, report overoptimistic beliefs about absolute performance in math. However, the bias is larger for men. The average guessed score in math is 7.7 for men (versus an average performance of 6.5 correct answers), and

[11] As intended, the average number of correct answers in Round 1 are also significantly lower for the harder versions of the quizzes in both domains, for both genders. Also note that the share of students who perform in the top-4 compared to the reference group is generally quite a bit lower than 40%. This is largely due to an unexpectedly high-performing reference group. The reference group students who were recruited prior to the roll-out were generally high-ability, recruited from honors classes. Since this impacts both genders equally, this should have no implications for our results.

6.4 for women (versus an actual performance of 5.7). Panel A of Figure D.3 shows that the distribution of overestimation of absolute performance in math is significantly different for men and women (p- value =0.000, based on a Kolmogorov-Smirnov (K-S) test).

Turning to verbal, the average guessed score is 7.6 for men (versus an average actual performance of 7.2), and 7.1 for women (versus an average actual performance of 6.9). Thus, the average bias in beliefs about absolute performance in verbal is smaller when compared to math. Panel B of Figure D.3 shows that the distributions do not differ significantly by gender (p-value of a K-S test=0.417). Panel B of Table 3.1 also shows that men report a higher average confidence in their beliefs for math (assigning higher probability to their guessed score), but the pattern reverses in verbal.

Turning to beliefs about *relative* performance, Panel B of Table 3.1 shows that the average guessed rank is higher for men than for women in both domains. The average (male − female) gap in guessed rank in math is 1.5 ranks (65% of the underlying SD in the measure), and in verbal is 0.6 ranks (32% of the SD). Panels C and D of Figure D.3 show that both genders, on average, have overoptimistic beliefs about relative performance in both domains. The size of the bias seems to be larger for men, particularly in math: the p-value of a K-S test for the equality of the two distributions in panel C of Figure D.3 is 0.004, and in panel D is 0.01.[12]

Panel B of Table 3.1 shows that the mean subjective probability of ranking in the

[12] Figure D.4 also reaffirms these patterns. Conditional on perceived absolute score, men tend to report a higher rank belief in both domains, especially in math. That is, not only do men tend to have larger biases in beliefs about their own absolute performance but they also perceive the population distribution of performance to be lower than women do (and hence, conditional on a score belief, place themselves higher in the rank distribution). A similar finding is reported in Coffman *et al.* (2021).

top-4 in math is 53% for men, versus 33% for women (p-value for equality of gender = 0.000). The gap is still sizable but relatively smaller in verbal: 50% for men and 39% for women (p-value = 0.000). Based on the performance of the reference pool, the actual proportion of individuals in the top-4 in Math is 31% for men and 18% for women. The corresponding proportions in Verbal are 27% and 24%. In short, both genders overestimate absolute and relative performance in math, with the bias being larger for men. There is, in general, both less overconfidence, and less of a male − female gap, in verbal.

Initial choices show similar patterns by gender and domain. Table 3.1 shows that men are more willing to accept competition in math than women: men, on average, need to be compensated $2.85 for each math problem to enter the competitive payment scheme versus $3.38 for women (p-value = 0.000). The average WTA in verbal is $3, and does not differ by gender. Under the fixed payment scheme, 56% of men choose math, compared to only 40% of women (p-value = 0.000).[13]

Table D.5 documents the relationship between initial beliefs and choices. As expected, initial beliefs have a sizable and significant impact on choices, and have predictive power even conditioning on actual performance. More optimistic beliefs about performance in a given domain are positively related with willingness to compete (that is, a lower WTA) in that domain. More optimistic beliefs in the other domain lead to a lower willingness to compete in the domain, though the magnitude of the estimates is less than half of the impact of the beliefs in the same domain.[14]

[13] 52 (60)% of the women (men) who chose math made the right decision based on their actual performance in Round 2. 73 (72)% of the women (men) who chose verbal made the right decision. See Table D.4.

[14] Approximately two-thirds of willingness to compete decisions maximize expected payoffs (without factoring in risk preferences) given stated beliefs, with no large differences by gender or timing

We now turn to understanding how those beliefs and choices respond to feedback.

The Provision of Feedback

Table D.6 reports the percentages of participants that receive each possible feedback combination, separately by gender. Throughout this chapter, we refer to good news as receiving feedback that you performed better than a randomly-chosen member of the reference group. While there is selection into type of feedback received based upon performance, conditional on rank (included as a control here and throughout our specifications), assignment to good and bad news is random. In our sample, 50% of women and 38% of men receive bad news in both domains, and 12.5% of women and 21% of men receive good news in both domains. This good − bad imbalance is a result of our (unexpectedly) talented reference group.

It is also worth noting that the signal structure we use, while simple, is informative. Table D.7 provides examples of what the posterior should be under Bayesian updating for various prior beliefs and levels of uncertainty. Recall that higher ranks correspond to better relative performance. Take a participant who assesses her relative performance to be low, assigning a probability of 20% each to ranks 1-5. This participant who has a prior belief of mean rank of 3 and fairly high uncertainty should revise her belief upward to 4.0 under Bayesian updating upon being informed that her performance is better than that of a randomly chosen person in the reference group, and should revise her mean rank belief down to 2.71 upon receiving a negative signal. For those with more optimistic priors about performance, the asymmetry of the Bayesian response is reversed. For instance, a respondent who has a prior belief of mean rank of 8 and fairly high uncertainty should revise her belief upward to just 8.29 after seeing good news, but downward to 7.00 after seeing bad news.

of choice.

As a first pass, Table 3.2 presents the rates at which participants adjust their beliefs of their own rank up, down, or not at all after receiving feedback. The table splits the data by type of feedback received and timing.[15] On average, participants respond to feedback in the direction we would expect. Immediately after receiving feedback, less than 10% of participants adjust their beliefs in the “wrong" direction (upward after bad news, or downward after good).[16] By one week later, this proportion grows to approximately 19%, suggesting already changes in reactions to feedback over time.

3.4.2 The Evolution of Beliefs and Choices Over Time

We start our main presentation of results by looking at individual level changes in beliefs and choices over time.

Individual Level Changes in Beliefs and Choices

We take as our starting point a model that will allow us to assess the direction and magnitude of shifts in beliefs and choices in response to feedback. We predict an individual's decision in a domain (either their belief or their choice) from when the decision was made: initially, immediately after feedback, or one week later. And, we account for whether the individual received good or bad news. We do so controlling

[15] Table D.8 provides these same statistics further split by gender and domain. The patterns are similar across domain and gender.

[16] Consistent with this pattern, we see that beliefs become more accurate after feedback on average: the mean squared error (MSE) in expected rank among treated participants falls from 11.3 initially to 8.8 and 9.0 immediately and one week after, respectively ($p < 0.001$ for each when compared to initial MSE). Mean squared error among our control participants is 11.7 initially and 11.6 one week later.

for a vector of individual controls, including their performance.[17]

We include seven dummies in the model to capture the various types of decisions we observe: Initial Decisions of Good News recipients, Immediate Decisions of Good News recipients, Week After Decisions of Good News recipients, Initial Decisions of Bad News recipients, Immediate Decisions of Bad News recipients, Week After Decisions of Bad News recipients, and Week After Decisions of the Control Group. The omitted category here is Initial Decisions of the Control Group; coefficients on each of the timing-feedback dummies should be interpreted as differences from Initial Decisions of the Control group.[18]

Formally, our model is:

$$\begin{aligned} O_{iDt} = &\beta_0 + \beta_1 \textit{Initial Good News}_{iDt} + \beta_2 \textit{Immediate Good News}_{iDt} \\ &+ \beta_3 \textit{Week After Good News}_{iDt} + \beta_4 \textit{Initial Bad News}_{iDt} \\ &+ \beta_5 \textit{Immediate Bad News}_{iDt} + \beta_6 \textit{Week After Bad News}_{iDt} \\ &+ \beta_7 \textit{Week After Control}_{iDt} + \mathbf{Y}_i + \mathbf{X}_i + \epsilon_{iDt}, \end{aligned} \tag{3.1}$$

where $D \in \{\text{Verbal, Math}\}$, O_{iDt} is a measure of beliefs or -WTA for participant i in domain D at stage $t \in \{\text{Initial, Immediate, Week After}\}$.[19]

[17] We control for performance and/or rank linearly in the regression analyses reported in the text. Our qualitative conclusions are unchanged if we instead use performance/rank fixed effects.

[18] We had not planned at the time of our pre-registration on including the control group in much of our analysis. But, after looking at the data and constructing a plan for making sense of it, it became clear that the control group provided a useful point of reference for interpreting the magnitudes of reactions to good and bad news. In our regression analysis in Table 3.4, we omit the control group to focus on whether there is differential updating by gender, timing, and type of news.

[19] Note that 79% of individuals make monotonic choices in every price list they see. In Appendix Table D.9, we show that our results are quite similar even when we restrict attention to only those participants who are always monotonic.

For both the beliefs and WTA measures, we use standardized measures, where higher numbers indicate better believed performance or more willingness to compete.[20] $\mathbf{Y}_i$ is a set of performance controls: the scores and rank of participant i in both domains and an indicator variable equal to one if the participant got the hard version of the tests in the first round. We use D to denote an observation associated with a given domain, and $-D$ to denote the other domain. $\mathbf{X}_i$ includes controls for family income, indicators for each parent attending college, a nonwhite indicator, ACT scores, high school rank, indicator for attending high school in the U.S., honors student indicator, school year (freshman, sophomore, junior or senior), a measure of risk aversion and an Immediate group indicator. It also includes an indicator for a female participant, and an indicator for whether the observation comes from math. We cluster standard errors at the individual level in each of these specifications.[21]

Figure 3.3 presents the results, plotting the coefficients of interest.[22] In particular,

[20] For the beliefs measure in the table, we create an aggregate measure that averages over beliefs of absolute performance, beliefs of rank, and beliefs of placing in the top 40 percent of performers. We use these measures to generate a standardized belief measure by domain with mean 0 and standard deviation of 1. At the baseline, the mean belief in math is 0.38 of a standard deviation for men and -0.26 for women (p-value$<$0.01). In verbal, the mean aggregate measure of beliefs is 0.22 of a standard deviation for men and -0.14 for women (p-value$<$ 0.01).

[21] We can also analyze the choice participants make about whether to choose a piece-rate in math over a piece-rate in verbal, though it requires a slight modification to the empirical approach (in particular, the two gender gaps, male-female and congruence, are indistinguishable) and the mechanism through which feedback impacts the choice is slightly less clear. To streamline presentation, we defer these results to Appendix E. Our results are qualitatively similar. In particular, we find that a significant immediate impact of feedback on individual choices, fading of that feedback over time, less fading of bad news for women compared to men, and limited impact of feedback on the gender gap.

[22] See columns (1) and (3) of Table D.10 for the regression estimates that produces this figure.

we normalize the omitted category (Initial Decisions of the Control Group) to 0 and plot the seven other coefficients relative to that baseline. Panel A considers standardized beliefs and Panel B considers standardized WTA.[23] Note that we re-sign WTA so that, in each panel, upward movement reflects more optimistic behavior – more positive beliefs about oneself or more willingness to compete.

We start by discussing beliefs. First, we should acknowledge the surprising difference in beliefs between individuals who *later* receive good or bad news. This indicates that, *prior* to the receipt of feedback, people assigned to receive good news had more optimistic beliefs than those assigned to bad news, conditional on true rank. This could reflect that individuals at the extreme ranks (1 or 10) are selected into good and bad feedback, i.e., a person with the top rank can only find out she performed better than a randomly-selected peer. We will account for this initial imbalance in our analysis going forward. In particular, decisions immediately after feedback or one week later should be compared to initial decisions, taking into account potentially different starting places for different types of news.

Turning our attention to the results of interest, we see that, consistent with Table D.8, beliefs on average move in the expected direction after feedback. Immediately after the receipt of good news, beliefs are 0.34 SDs more optimistic than initially.[24]

Note that while equation (1) controls for performance in the other domain, it does not control for specific feedback received in that domain. Column (2) of Table D.10 shows that controlling for specific feedback has no impact on the estimates for beliefs; column (4) shows that controlling for feedback in the other domain yields the same qualitative conclusions for the impacts on the WTA (not surprisingly, receiving bad news in the other domain makes individuals more willing to compete in a given domain).

[23] If we analyze instead either beliefs of absolute score or believed rank in isolation, our results throughout this chapter look quite similar. Results available upon request.

[24] This can be observed by subtracting the coefficient on Initial Good News from the coefficient

Immediately after the receipt of bad news, beliefs are 0.21 SDs more pessimistic than initially. Thus, we see a sizable and significant immediate response to both types of feedback. By one week later, this impact has faded significantly. Beliefs one week after good news are only 0.23 SDs more optimistic than initial beliefs (0.11 SDs less optimistic than beliefs immediately after good news, $p<0.001$). There is less fading after bad news: beliefs one week after bad news are 0.17 SDs more pessimistic than initially (just 0.04 SDs more optimistic than beliefs immediately after bad news, $p=0.02$).[25] Both positive and negative feedback have a strong initial impact on beliefs; while some of the impact of good news fades over time, there is still a sizable impact of feedback one week later.

Patterns for WTA look similar to the patterns for beliefs. Individuals who receive good news are significantly more willing to compete immediately after feedback than prior to feedback (by 0.26 SDs, $p<0.001$); and, individuals who receive bad news are significantly less willing to compete immediately after feedback than prior to feedback (by 0.19 SDs, $p<0.001$). By one week later, again some of the impact of the good news has faded: good news recipients are just 0.17 SDs more willing to compete than they were initially ($p=0.005$ when we compare week after to immediately after). The bad news impact also fades, with individuals 0.11 SDs less willing to compete one week later compared to initially ($p=0.005$ when we compare week after to immediately after).[26]

Finally, we point out the interesting action in the Control group. Despite receiving

on Immediate Good News.

[25] In fact, the change in beliefs between immediately after feedback and one week later is significantly greater for good news than bad, $p=0.002$.

[26] We cannot reject that the extent of fading between immediately after feedback and one week later is the same after good and bad news for choices, $p=0.67$.

no feedback on performance, individuals in the Control group become more optimistic over time: they have significantly more optimistic beliefs about themselves one week later than initially (by 0.12 SDs, p<0.001) and are more willing to compete (by 0.06 SDs, p=0.07). In our setting, no news seems to be good news.

This trend in the Control group also has implications for how we think about asymmetry in reactions to good and bad news. One could simply compare the absolute value of the change from initial to week after beliefs for good news and bad news. But, an alternative - and perhaps more appropriate - way is to ask whether the changes in response to news, *relative to the changes in the Control group*, are larger for good versus bad news. These two methods will not necessarily produce the same answer, given the positive trend in the Control group. Consider Panel A on Beliefs. The absolute change for Good News is larger (over one week) than the change for Bad News: 0.23 SDs for change in good news versus 0.17 SDs for change in bad news, p=0.002. But, when we look at responses relative to the Control group, it is the reaction to Bad News that is larger: beliefs after good news grow by just 0.11 SDs more than they do in the control group (p<0.001), while beliefs after bad news fall by 0.29 SDs more than they do in the control group (p<0.001). Relative to the Control group, it is bad news that is having the larger impact on decisions over time.

We have documented that our feedback has a significant overall impact on beliefs and choices. On average, individuals become significantly more optimistic and more willing to compete after receiving good news, though these effects get weaker over time. Bad news makes participants significantly less optimistic and less willing to compete, and these effects seem to be rather persistent one week later.

Differences by Gender and Stereotype

A natural next question is whether these patterns vary by gender or the gender congruence of the domain. To explore this, we adapt our model to estimate these reactions either (i) separately for men and women, or (ii) separately for gender congruent and incongruent domains. We take the model from equation (1) but expand it, first, to include a full set of dummies for each gender-news-timing combination (Initial Decisions of Women who Receive Good News, Initial Decisions of Men who Receive Good News, ...). Second, and separately, we expand the model to include a full set of dummies for each congruence-news-timing combination (Initial Decisions for People who Receive Good News in a Congruent Domain, Initial Decisions for People who Receive Good News in a Incongruent Domain, ...). We define congruent as participant i is a woman and the domain D is verbal or when i is a man and D is math.[27]

Figure 3.4 presents the results for gender.[28] For men, we normalize the Initial Decisions of Men in the Control Group to 0, and simply plot the coefficients on the other male dummies. Each of these plotted points for men can be interpreted as differences from the Initial Decisions of Men in the Control Group. To facilitate comparisons of trends over time, we make the choice to also normalize the Initial Decisions of Women in the Control Group to 0, and adjust all of the coefficients on the female dummies accordingly. In this way, the plotted points for each coefficient associated with women can be interpreted as differences from the Initial Decisions

[27] In columns 3 and 4 of Table D.11, we show that these results are quite similar if we instead define congruent according to the participant's own stated beliefs, assigning a 1 for each domain the participant indicated they believed their own gender had an advantage in. Our definition of domain congruence matches an individual's stated beliefs in more than 60% of cases.

[28] See Table D.12 for the regression estimates that produces this figure.

of Women in the Control Group. While this does make *trends over time* easier to compare and interpret across gender, we should point out that it differences out the initial male − female gap. That is, this figure completely hides the fact that, conditional on performance, women have significantly less optimistic beliefs and are significantly less willing to compete initially (and subsequently). Our focus here is on how men's and women's decisions evolve over time; we will return to the implications of these patterns for gaps later in our results section.

We start by discussing beliefs, that are shown in Panel A of Figure 3.4. First, we point out that the positive trend in the control group that we observed in Figure 3.3 holds for both men and women, who both grow approximately 0.11 SDs more optimistic over time in the control group ($p<0.001$ for both). After good news, women adjust their beliefs up by 0.40 SDs immediately; beliefs one week after good news have faded, but are still 0.30 SDs more optimistic than initially ($p<0.001$ both when comparing week after to initial and when comparing week after to immediate). Women's absolute adjustments after bad news (relative to good) are smaller, but fade less. After bad news, women's beliefs are 0.17 SDs more pessimistic immediately and this is essentially unchanged one week later ($p=0.67$ comparing week after beliefs to immediate beliefs after bad news). The impact of bad news fades less for women than the impact of good news ($p=0.01$).

Men, on the other hand, see significant fading of reactions after both good and bad news. Men adjust their beliefs up by 0.26 SDs immediately after good news, and week after beliefs are 0.15 SDs more optimistic than initially ($p=0.001$ comparing week after to immediate). The pattern for bad news is pretty symmetric. They adjust their beliefs down 0.28 SDs immediately after bad news. But, by one week later, beliefs are just 0.20 SDs more pessimistic than initially ($p=0.011$ when comparing week after to immediate). The amount of fading is no different after good or bad news for men (p=0.55).

The punchlines are quite similar when looking at choices (Panel B). In particular, we see significant and sizable reactions to good and bad news, for both men and women. And, as with beliefs, men and women show different patterns in terms of what type of news fades. For women, the impact of good news fades significantly (by 0.12 SDs, p=0.006); the impact of bad news does not (0.06 SDs, p=0.13; p=0.23 on the difference-in-difference). For men, it is the impact of good news that does not fade significantly (0.06 SDs, p=0.35), and the impact of bad news that does fade (by 0.10 SDs, p=0.03; p=0.43 on the difference-in-difference).

Summarizing the evidence for gender, we see significant, sizable reactions to good and bad news for both men and women. We see some evidence that the impact of bad news persists more than the impact of good news for women. This is not the case for men.

We next turn to the result for congruence in Figure 3.5.[29] For incongruent observations, we normalize the Initial Decisions for Incongruent Domains in the Control Group to 0, and simply plot the coefficients on the other incongruent dummies. Each of these plotted points can be interpreted as differences from the Initial Decisions in Incongruent Domains in the Control Group. Again, to facilitate comparisons of trends over time, we make the choice to also normalize the Initial Decisions in Congruent Domains in the Control Group to 0, and adjust all of the coefficients on the congruent dummies accordingly. In this way, the plotted points for congruent domains can be interpreted as differences from the Initial Decisions in Congruent Domains in the Control Group. We offer the same caveat as we did for gender: this presentation differences out the initial congruent − incongruent gap. In fact, conditional on performance, individuals are significantly more optimistic and more willing to compete

[29] Table D.11 shows the estimates that produces this figure.

in congruent domains than incongruent domains. But, we leave our consideration of these gaps to our later discussion. For now, we focus on how decisions evolve over time within both congruent and incongruent domains.

Panel A considers beliefs. We see a large degree of similarity across congruent and incongruent domains. After receiving good news in an incongruent domain, individuals adjust their beliefs up by 0.33 SDs immediately after feedback; beliefs one week later have fallen back by 0.09 SDs (p<0.01 when comparing week after and immediately after). Reactions after bad news are also sizable: after receiving bad news in an incongruent domain, individuals adjust down by 0.26 SDs immediately. This bad news reaction does not fade, with beliefs remaining 0.24 SDs more pessimistic than initially one week later (p=0.36 when comparing week after and immediately after). When we turn our attention to congruent domains, we see a similar set of immediate reactions: individuals adjust up by 0.34 SDs in response to good news, and down by 0.26 SDs after bad news. Again, reactions to good news fade by approximately 0.11 SDs over the course of the week (p<0.001 comparing week after to immediately after). But, unlike in the case of incongruent domains, for congruent domains, the bad news reactions fade as well, rising by 0.06 SDs over the course of the week (p=0.02 when comparing week after to immediately after).

Panel B presents the results for willingness to compete, where the patterns are largely similar. In particular, reactions to good news are large for both domain types - roughly 1/4 of a SD – and fade over time (by approximately 0.1 SDs, though the fading is not significant for incongruent domains, p=0.12). Individuals who receive bad news in an incongruent domain adjust their willingness to compete down by 0.16 SDs immediately, and the effect is similar one week later (remaining at 0.12 SDs, p=0.32 on the comparison). Individuals who receive bad news in a congruent domain adjust down by 0.23 SDs immediately, before bouncing back up by 0.12 SDs (p=0.02

on the comparison of week after and immediately after).

Thus, both for women and for individuals in incongruent domains, we see three consistent patterns: (i) immediate good news reactions are larger than bad news reactions, (ii) good news reactions fade over the course of a week, and (iii) bad news reactions do not fade.

As we mentioned, this analysis is helpful in considering within-gender or within-domain-type trends over time. However, making comparisons across gender or congruence in this setting is harder. In particular, men and women (and individuals in gender-congruent versus gender-incongruent domains) begin with different initial beliefs and choices, conditional on performance. There may be more “room" for upward or downward adjustment among some of these groups given their starting points. Thus, we will return to this issue with a specific focus on gaps and comparisons across gender and congruence later in our results section. There, we will see how these individual trends over time map into gender gaps.

While not our primary focus, one could also use our data to consider whether participants update their beliefs in a manner consistent with Bayes rule. Since we elicit the full subjective distribution of prior rank beliefs, we can indeed construct a Bayesian benchmark for each individual's beliefs about her rank, given her prior belief distribution and the feedback she receives. In Appendix Table D.13, we regress the individual's posterior belief of her rank onto the Bayesian posterior, a dummy for having received good news, the interaction of the two, and a constant term, separately by gender and domain, both immediately and a week later. It is worth noting that this specification does not control for performance (since the regression interpretation would then be unclear), and hence the feedback is not randomly assigned. Thus, one should be cautious in interpreting these results beyond a within-specification test of the Bayesian model. Updating that is fully consistent with Bayesian updating would imply that the constant term should be zero and the estimate on the Bayesian posterior should be one. That is not the case. Consistent with existing literature (Benjamin, 2019), we see that updating tends to be conservative (that is, information is more likely to be discounted relative to a Bayesian benchmark), both immediately after feedback as well as a week later. There seems to be more conservativeness after bad news than good, particularly for men and for congruent domains. No gender overreacts to certain kinds of news in either domain. However, we again caution that feedback cannot be interpreted as randomly assigned in these specifications.

The Role of Recall

Figure 3.3 shows a somewhat fading impact of feedback over time. Both for beliefs and choices, decisions one week later (relative to immediately after feedback) seem to fall back closer to baseline. One natural candidate explanation for this pattern is forgetting. Could it be that participants simply forget the feedback they received? Our two-session design allows us to consider how well participants recall feedback one week later (at the end of Session 2). In addition to overall rates of forgetting, we can explore interesting heterogeneity. Does the accuracy of recall vary with gender, or the type of feedback received (good, bad)? Consistent with a motivated reasoning story as in Zimmermann (2020), Chew *et al.* (2020), and Huffman *et al.* (2019), are individuals more likely to remember good news than bad?

Overall, the rate of accurate recall is high: 88% of feedback received is accurately recalled.[30] Figure 3.6 reports the rate of accurate recall by type of feedback. It is clear that participants who received mixed feedback – good in one domain, bad in the other – are less likely to accurately recall their feedback.

In Column (1) of Table 3.3, we regress a dummy for accurately recalling the feed-

[30] This high recall rate also suggests that our participants were attentive on average.

back received in a domain onto indicators for the participant's gender, whether the domain is gender-congruent, and whether the feedback was good news. We control for performance, including rank, as well as our standard demographic controls. We include a dummy variable for whether or not the participant was assigned to the Immediate group, the group that is asked to report beliefs and choices both immediately after the provision of feedback and one week later; this is to allow for the possibility that these individuals, having been asked to immediately react to it, may be more likely to recall this feedback one week later. In column (2), we also control for the type of feedback received in the opposite domain, and the interaction of the feedback from both domains to pick up potential confusion of the two pieces of feedback. In columns (3), (4) and (5) we include interactions of good news with gender, congruence of the domain and Immediate to analyze potential heterogeneous effects. We use only the sample that receives feedback, omitting the Control group, and we cluster errors at the individual level.

Most strikingly, we see that individuals are significantly more likely to recall bad news than good: column (1) shows that individuals are 9pp less likely to accurately recall good news compared to bad. In column (2), we see that this is mostly driven by people who received mixed feedback. Participants are more likely to recall one piece of good news correctly when they also received good news in the other domain ($p<0.01$).

We also find that women are 6-7pp more likely to accurately recall feedback than men. This male $-$ female gap is indistinguishable for good and bad news (column 3). Overall, the gender congruence of the domain has no significant predictive power for the accuracy of recall; while good news is directionally more likely to be recalled when it is received in a congruent domain compared to an incongruent domain, this effect is not large or statistically significant (Column 4). We do not find evidence that people in the Immediate group are more likely to recall their feedback, nor is it the case that being assigned to the Immediate treatment changes the amount of good-bad asymmetry in recall (column 5).[31]

In Appendix Figures D.5 and D.6, we show that the patterns we documented in Figures 3.4 and 3.5 do not appear driven by forgetting of feedback. In particular, if we reproduce these figures restricting attention to only those observations for which the feedback was accurately recalled, the patterns look quite similar. Thus, the fading of feedback over time and the greater persistence of bad news compared to good does not seem entirely explained by patterns of recall.

3.4.3 The Evolution of Gaps Over Time

In this section, we consider the implications of our results for gender gaps in beliefs and choices. In particular, we document the male − female gap and the congruence gap at each point in time, and ask whether feedback helps to reduce these gaps.

[31] It is worth noting that while Zimmermann (2020) finds evidence of motivated recall and updating when participants are surveyed one month after feedback, the good-bad asymmetry is reduced when participants are surveyed immediately after feedback, when they are given large incentives for accuracy, or when they know in advance they will be rewarded for accurate beliefs. Our setting, with a shorter delay and in which accurate beliefs can help to improve payments in Round 2, may not provide the type of wiggle room needed for this type of motivated reasoning to occur.

Plotting Gender Gaps Over Time

We begin with analysis of the male − female gap across the three points, estimating the following model:

$$\begin{aligned}
O_{iDt} = & \beta_0 + \beta_1 \mathit{Immediate}_{iDt} + \beta_2 \mathit{Week\ After}_{iDt} + \beta_3 \mathit{Female}_i \times \mathit{Initial}_{iDt} \\
& + \beta_4 \mathit{Female}_i \times \mathit{Immediate}_{iDt} + \beta_5 \mathit{Female}_i \times \mathit{Week\ After}_{iDt} \\
& + \beta_6 \mathit{Initial\ Control\ Group}_{iDt} + \beta_7 \mathit{Week\ After\ Control\ Group}_{iDt} \\
& + \beta_8 \mathit{Female}_i \times \mathit{Initial\ Control\ Group}_{iDt} \\
& + \beta_9 \mathit{Female}_i \times \mathit{Week\ After\ Control\ Group}_{iDt} + \mathbf{Y}_i + \mathbf{X}_i + \epsilon_{iDt},
\end{aligned} \tag{3.2}$$

where $D \in$ {Verbal, Math}, O_{iDt} is a measure of beliefs or -WTA for participant i in domain D at stage $t \in$ {Initial, Immediate, Week After}. As before, for both the beliefs and WTA measures, we use the standardized measures, where higher numbers indicate better believed performance or greater willingness to compete. The estimates in the specification are relative to the omitted group of Initial treated male respondents. The controls are the same as in equation (1); one difference is that since we are now also interested in the congruence gap, instead of an indicator for math, we use an indicator for whether the observation comes from a gender-congruent domain. This variable takes value one when domain D is congruent with i's gender, that is, when participant i is a woman and the domain D is verbal or when i is a man and D is math.

Since the vector $\mathbf{X}_i$ includes an indicator for whether the participant is female, the parameters β_3, β_4, and β_5 show the male − female gap in the outcome at each stage of the experiment. These estimates are plotted in Figure 3.7.[32] Importantly, these are gaps controlling for performance. After completing the quizzes but prior

[32] See Table D.14 for the corresponding regression estimates.

to receiving feedback, we observe a significant male − female gap in believed performance of 0.35 standard deviations. The provision of feedback significantly reduces this gender gap (to about 0.26 SDs, p=0.03 comparing immediate gap to initial gap). One week later, part of the impact has dissipated: the male − female gap in beliefs moves directionally closer to its starting point, at 0.30 SDs. This final gender gap is statistically indistinguishable from the gap immediately after feedback (p=0.23), and significantly smaller than the starting gap (p=0.04).

The second panel considers willingness to compete. The initial male − female gap in the measure is approximately 0.14 SDs (that is, women have to be compensated about 0.14 of a standard deviation more to accept the competitive pay scheme). As in the left panel for beliefs, we again see an inverse u-shaped pattern: feedback reduces the immediate male − female gap to 0.09 SDs (though the estimate does not statistically differ from the initial estimate, p-value= 0.39). However, a week later, the male − female gap is back at its starting point, at 0.15 SDs (p=0.26 comparing final and immediate gaps).

We see a very similar pattern of results when we consider the gender congruence gaps. We adapt Equation (2), replacing "female" with "congruent domain."[33] Figure 3.8 plots the congruence gap across the three points in time.[34] A positive congruence gap indicates that individuals have more optimistic beliefs and are more willing to compete in a domain that is congruent with their gender, controlling for measured performance.

The left panel shows that, initially, individuals are significantly more optimistic

[33] Note that we still include a female indicator in this model.

[34] See Table D.15 for the regression estimates that produces this figure. In columns 3 and 4, we show that these results are quite similar if we instead define congruent according to the participant's own stated beliefs.

about their performance in gender-congruent domains: conditional on actual performance, individuals are 0.16 standard deviations more optimistic in the gender-congruent domain. Just as feedback reduced the male − female gap in beliefs, feedback directionally reduces the gender-congruence effect. The estimated impact of gender congruence falls to 0.13; this gap is quite similar one week later, with a final coefficient on gender congruence of approximately 0.14 SDs. None of these gaps are significantly different from each other. The message is largely the same in the right-hand side panel for -WTA. The initial congruence effect is 0.19 SDs. Immediately post-feedback, this drops to 0.17 standard deviations, before closing one week later back at 0.19 SDs. Again, none of these gaps are significantly different than each other.

Thus, the only gender gap that feedback has significantly reduced one week later is the male − female gap in beliefs. Even in that case, the gap remains sizable, having fallen from 0.35 SDs to 0.30 SDs.[35]

Gaps after Good and Bad News

We have documented that feedback is largely ineffective in reducing gender gaps. In this section, we ask, are good and bad news equally (in)effective in reducing gender gaps? That is, do gaps evolve differently among individuals who (exogenously) receive good versus bad news? To test this, we expand equation (2) to consider the potential

[35] In Table D.16, we show that our intervention also has a minimal overall impact on gender gaps in payoffs. In addition, Figure D.7 reports the realized expected payoffs as a percentage of the maximum achievable payoff at every point in time, split by gender. On average, the expected payoffs as a percentage of the maximum achievable payoff are similar by gender: 63% for males and 65% for females at the initial stage. Additionally, we see that receiving feedback does not get participants any closer to their maximum achievable earnings, not immediately after nor a week later.

for differential effects for good versus bad news. In particular, for the male − female gap, we have:

$$\begin{aligned}O_{iDt} =&\beta_0 + \beta_1 \textit{Immediate Good News}_{iDt} + \beta_2 \textit{Week After Good News}_{iDt}\\ &+ \beta_3 \textit{Female}_i \times \textit{Initial Good News}_{iDt} + \beta_4 \textit{Female}_i \times \textit{Immediate Good News}_{iDt}\\ &+ \beta_5 \textit{Female}_i \times \textit{Week After Good News}_{iDt} + \beta_6 \textit{Initial Bad News}_{iDt}\\ &+ \beta_7 \textit{Immediate Bad News}_{iDt} + \beta_8 \textit{Week After Bad News}_{iDt}\\ &+ \beta_9 \textit{Female}_i \times \textit{Initial Bad News}_{iDt} + \beta_{10} \textit{Female}_i \times \textit{Immediate Bad News}_{iDt}\\ &+ \beta_{11} \textit{Female}_i \times \textit{Week After Bad News}_{iDt} + \beta_{12} \textit{Initial Control Group}_{iDt}\\ &+ \beta_{13} \textit{Week After Control Group}_{iDt} + \beta_{14} \textit{Female}_i \times \textit{Initial Control Group}_{iDt}\\ &+ \beta_{15} \textit{Female}_i \times \textit{Week After Control Group}_{iDt} + \mathbf{Y}_i + \mathbf{X}_i + \epsilon_{iDt}.\end{aligned} \tag{3.3}$$

where the controls are exactly as in equation (2), and the omitted category is the Initial treated male respondents. As before, we control for actual performance, and so assignment to good or bad news is random. The parameters β_3 and β_5, for example, reflect the male − female gap at the initial stage of the experiment for individuals who go on to receive good news and bad news, respectively.

The parameters of interest are β_3, β_4, β_5, β_9, β_{10}, and β_{11}. These are plotted in Figure 3.9. We see that the gender gaps in both beliefs and choices (reassuringly) start out quite similar across the groups that go on to receive good versus bad news. For choices, the male − female gaps for good and bad news also evolve similarly over time, both first shrinking somewhat immediately in response to feedback, before inching back towards their initial starting points. But, for beliefs, we observe a divergence. While both good and bad feedback shrink the gender gap immediately, this is no longer the case one week later. The male − female beliefs gap after good news does not bounce back towards its starting point. But, the gap after bad news does. As a result, the final male − female gap in beliefs is significantly larger after bad news than good (p=0.005).

We consider congruence gaps in Figure 3.10. Again, the initial congruence gaps are similar across the groups who go on to receive good and bad news, as expected. For beliefs, the congruence gaps for good compared to bad news are indistinguishable at any of the three points in time, and the gaps do not change significantly differently over time. But, for choices, bad news seems more problematic. While the congruence gap for good news directionally falls at each point in time, the congruence gap for bad news falls initially before bouncing back strongly. Again, the result is that the final congruence gap for choices is significantly larger after bad news than good (p=0.034).

Our specifications so far have focused on the evolution of both individual beliefs and choices over time, and gaps in beliefs and choices over time. In producing these estimates, we have been careful to account for performance. Our analysis, particularly that in Figures 3.9 and 3.10, shows that there are differences (both male − female and congruent − incongruent differences) in how two individuals with the same performance and who receive the same feedback update their beliefs and choices over time. These differences seem to be starker for bad news recipients, particularly in the longer run (one week later).

In this final section, we push this analysis one step farther, asking whether there are gender differences in beliefs and choices across individuals with the same performance, who receive the same feedback, and also have the same *initial decisions.* Our focus will be on understanding the explanatory power of initial decisions, prior to feedback, in predicting beliefs and choices immediately and one week after feedback. We do this first for beliefs, and then for choices. This allows us to ask how much the initial gender differences in beliefs and choices matter for persistence of the gaps. It could be the case that men and women actually respond quite similarly to feedback, conditional on initial beliefs and choices, but that initial beliefs and choices are very different. These different starting points may matter a lot for how individuals respond to feedback. In that case, it is the stickiness of initial beliefs and decisions that fuels persistence. Alternatively, it could be the case that, given the same initial beliefs or choices, men and women respond differently to feedback in ways that further perpetuate initial gaps.

We adjust our empirical approach to focus on estimating gaps for each particular point in time, initially, immediately after feedback, and one week later. For immediately after feedback and one week after feedback, we estimate this equation first without the initial belief/decision, and then with the initial belief/decision:

$$\begin{aligned} O_{iD} =& \beta_0 + \beta_1 \mathit{Bad\ News}_{iD} + \beta_2 \mathit{Bad\ News}_{iD} \times \mathit{Female}_i + \beta_3 \mathit{Good\ News}_{iD} \times \mathit{Female}_i \\ &+ \beta_4 \mathit{Bad\ News}_{iD} \times \mathit{Congruent}_{iD} + \beta_5 \mathit{Good\ News}_{iD} \times \mathit{Congruent} \\ &+ \beta_6 \mathit{Prior\ O}_{iD} + \mathbf{Y}_i + \mathbf{X}_i + \epsilon_{iD}, \end{aligned} \tag{3.4}$$

where $D \in \{\text{Verbal, Math}\}$, O_{iD} is a measure of beliefs or -WTA for participant i in domain D, and $\mathit{Prior}\ O_{iD}$ is the initial outcome (belief or -WTA) when the model is estimated using the immediate or week after decisions. $\mathit{Congruent}_{iD}$ is a dummy that equals 1 when the participant i is female and the domain D is verbal, or when i is male and the domain is math, and zero otherwise.[36] The variables in $\mathbf{Y}_i$ and $\mathbf{X}_i$ are the same as in equation (2) except that they no longer include an indicator for whether the respondent is a female and whether the domain is gender-congruent (since those terms are now shown explicitly).

Panel A of Table 3.4 presents the results for beliefs. Column (1) estimates the ini-

[36] In this case, the omitted category is males who receive good news in the gender-incongruent domain, i.e., Verbal.

tial gender gaps conditional on performance and news received (the omitted category is good news for males in the incongruent domain). In column (1), the insignificant estimate on *Bad News*$_{iD}$ indicates that males who go on to receive bad news have similar initial beliefs as their male counterparts who go on to receive good news (in the incongruent domain). Females, on the other hand, have significantly lower beliefs than their male counterparts. Initial beliefs in congruent categories are more favorable/higher. Moving from Column (1) to Column (2) and then to Column (4) shows how the gaps evolve over time, first immediately after feedback and then one week later later (that is, the Bad News*Female and Good News*Female terms). These results largely echo the observations of Figures 3.9 and 3.10. Our focus in this analysis to ask what happens to these estimated gaps once we account for initial beliefs. When we compare Column (2) to Column (3), we ask, how much of the residual gender gaps immediately after feedback can be explained by differences in prior beliefs? We see that while gender and congruence gaps after *good news* are still sizable immediately after feedback (a 0.21 SD male-female gap and a 0.14 SD congruence gap, as shown in Column 2), they are entirely explained by differences in initial beliefs (coefficients on Good News*Female and Good News*Congruent of close to 0 in Column 3). This is not the case for the male − female gap after bad news. Even once we account for initial beliefs, we estimate that women's beliefs immediately after bad news are 0.08 SDs more pessimistic than men's (p=0.02). The message when comparing Columns (4) and (5) is similar. While there are significant gaps after good news one week after feedback, they are explained by differences in prior beliefs. In fact, conditional on initial beliefs and performance, women's beliefs after good news are actually more optimistic than men's one week later. But when we turn our attention to bad news, the residual male − female gap is not fully explained. Conditional on performance and prior beliefs, women's beliefs are 0.07 SDs more pessimistic than men's one week later (p=0.004). Both immediately and one week later, congruence gaps after bad news do seem to be fully explained by initial beliefs.

Panel B of Table 3.4 presents the same analysis for choices, predicting our standardized willingness to compete measure (-WTA). Keep in mind that here, we ask whether residual gaps in choices can be explained by differences in initial *choices*. We are not adding prior beliefs to model. Instead, we are asking whether two men and women who started in the same place, in terms of choices, look the same one week later.

Just as we saw with beliefs, we see that there are no significant gender gaps in choices after good news, once we account for initial decisions. Conditional on having the same performance, making the same initial choices, and receiving positive feedback, men and women are equally willing to compete one week later, and individuals are equally willing to compete across congruent and incongruent domains. After bad news, we do see differences. One week after receiving bad news, women are 0.19 SDs less willing to compete than men (p=0.0001, Column 4). Even once we condition on initial decisions, we continue to estimate that women are 0.08 SDs less willing to compete after bad news than men (p=0.041, Column 5). This is also true when we consider congruence gaps. One week after bad news, individuals are 0.23 SDs less willing to compete in incongruent domains compared to congruent domains - even conditional on having the same performance in each (p<0.01, Column 4). Again, controlling for initial decisions fails to close this gap. Even conditional on having the same performance and the same initial willingness to compete, individuals are 0.10 SDs less willing to compete in incongruent domains compared to congruent domains one week after feedback (p<0.01). These residual choices gaps are consistent with differential updating in response to bad news across men and women, and across congruent and incongruent domains.

For both beliefs and choices, our data show that gender gaps seem to persist after bad news. Table 3.4 highlights that this persistence is not fully explained by differences in initial decisions. Even conditional on having the same performance and making the same initial decisions, men and women seem to update their beliefs and choices differently in response to bad news.

In Table D.17, we show that these residual gaps (for beliefs) are also unexplained by a Bayesian model. In particular, we add to the specifications of Panel A of Table 3.4 the Bayesian predicted posterior as an additional explanatory variable (note that the dependent variable here is the expected rank, which is different from the standardized belief measure used in Table 3.4). If the residual gaps were consistent with Bayesian predictions, we would expect no significant gender differences after we include this predicted posterior as a control. Instead, we find that the inclusion of the Bayesian prediction has limited additional impact on the estimated gender gaps. We conclude that there are gender differences in how men and women update their beliefs in response to bad news in our environment, beyond what a Bayesian model would predict.

3.5 Conclusion

The potential of information provision for reducing gender gaps depends on how women and men respond to feedback. Prior literature primarily studies the role of information in static settings. However, many important contexts – education, for example – are dynamic in nature. Therefore, it is necessary to understand how beliefs and choices respond to feedback immediately after its provision and how this response might change over time. We explore the dynamics of belief updating over time, with an emphasis on understanding the role that gender and stereotypes play, and the impact on not only beliefs, but choices. In this chapter, we take an important step toward answering these questions in an experiment that is dynamic by design. We complement recent work on gender differences in choices after failure by designing an experiment that identifies underlying channels.

In line with existing literature that finds that information interventions can impact beliefs and behaviors, we find sizable immediate impacts of feedback on beliefs and choices (with impacts in the range of 0.2 - 0.35 standard deviations). While these impacts partly fade out a week later (and the fade out patterns depend on the type of news that is received), they remain economically and statistically significant.

Turning to gender gaps, we find that feedback reduces male $-$ female gaps in beliefs and choices immediately after feedback, but a week later part of this effect dissipates. Similarly, although feedback reduces the gender-congruence gap in beliefs and choices immediately after feedback, the gap reverts to its initial level after a week. Our design allows us to show that the persistence of gender gaps is not due to forgetting feedback or differential recall. Conditional on performance and initial decisions, we find that women and men update their beliefs and choices similarly in response to *positive* feedback. The same is not true for updating after bad news. One week after receiving negative feedback, women hold more pessimistic beliefs and are less willing to compete than men with the same performance and initial decisions. It is, however, worth nothing that both genders under-react to feedback relative to a Bayesian benchmark, regardless of the news type.

Beliefs and choices evolve differently for men and women after negative feedback. There seems to be a pull toward gender gaps, in the longer run, even conditional on starting point and feedback received. What drives this pull – be it cognitive or motivated biases, tastes, norms, or other forces – remains an important open question. However, our findings offer a cautionary note to the promise of one-time information interventions to address gender gaps. Repeated provision of feedback at higher frequencies may be more effective in eliminating biases and stereotypes, and should be explored in future work. Yet, the fact that we (and others) find significant gender biases in initial beliefs and choices, even in environments where individuals are likely to have received many signals in the past, suggests that even richer informational environments may fail to fully close gender gaps.

A major implication of our results is that prior beliefs/choices continue to be important in explaining the changes (or lack thereof) over time. Thus, a better understanding of how initial beliefs are formed, and how tastes for different domains emerge, is crucial for understanding decision-making at the individual level as well as shedding light on the stubbornness of gender gaps.

It is also worth noting that we do not find evidence of motivated memory. Participants in our setting are more likely to recall negative feedback than positive feedback. And, the impact of good news seems to fade more than the impact of bad news. This is somewhat inconsistent with recent papers that have either found higher recall of positive feedback or positively biased beliefs about past performance. It could be that, in our context, accurate beliefs can help improve payoffs, and this mitigates the role of motivated memory or beliefs. In any case, more work is needed to understand when such biases may emerge.

Figures and Tables

Figure 3.1: Experimental Design

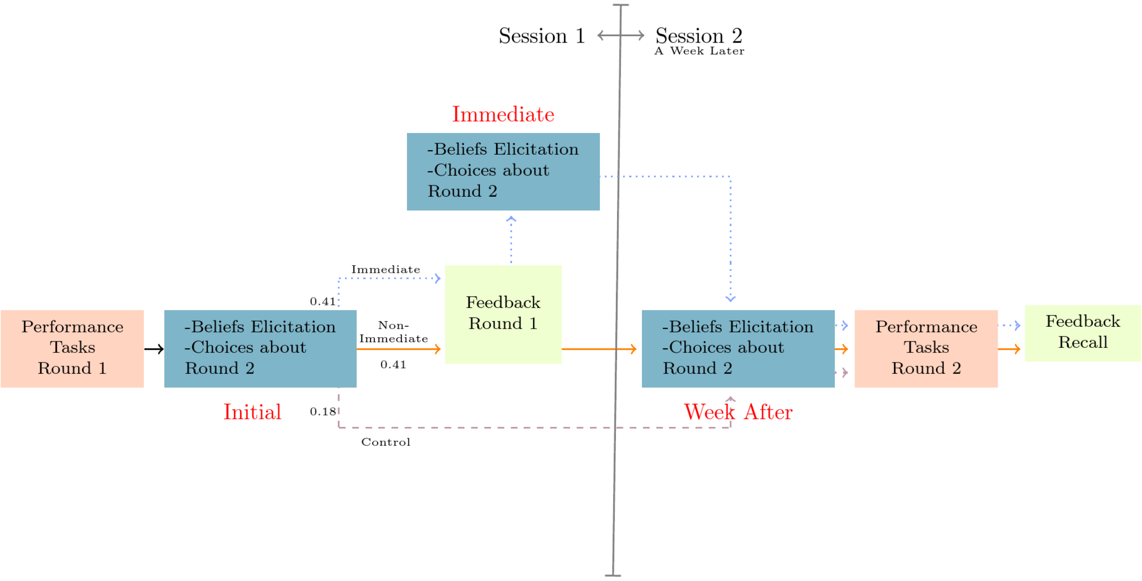
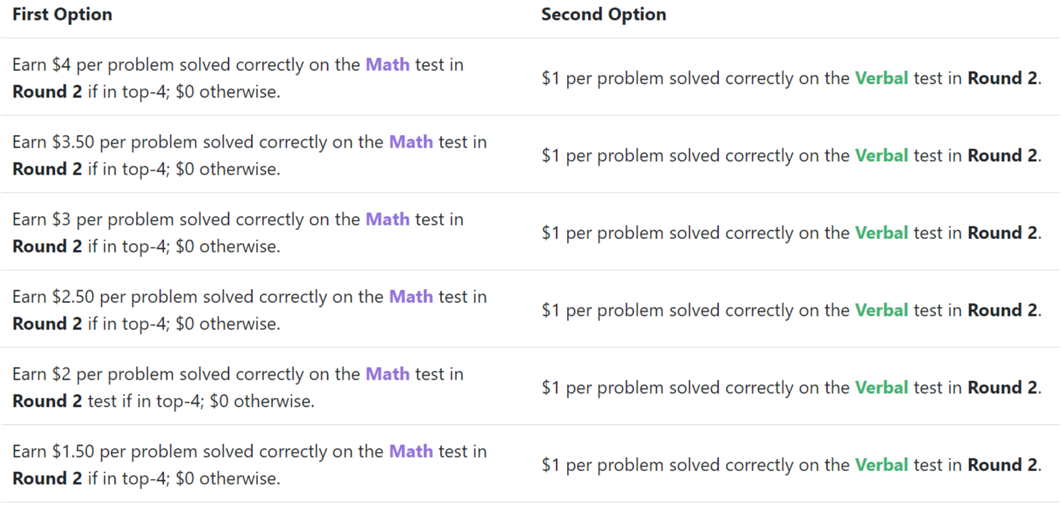
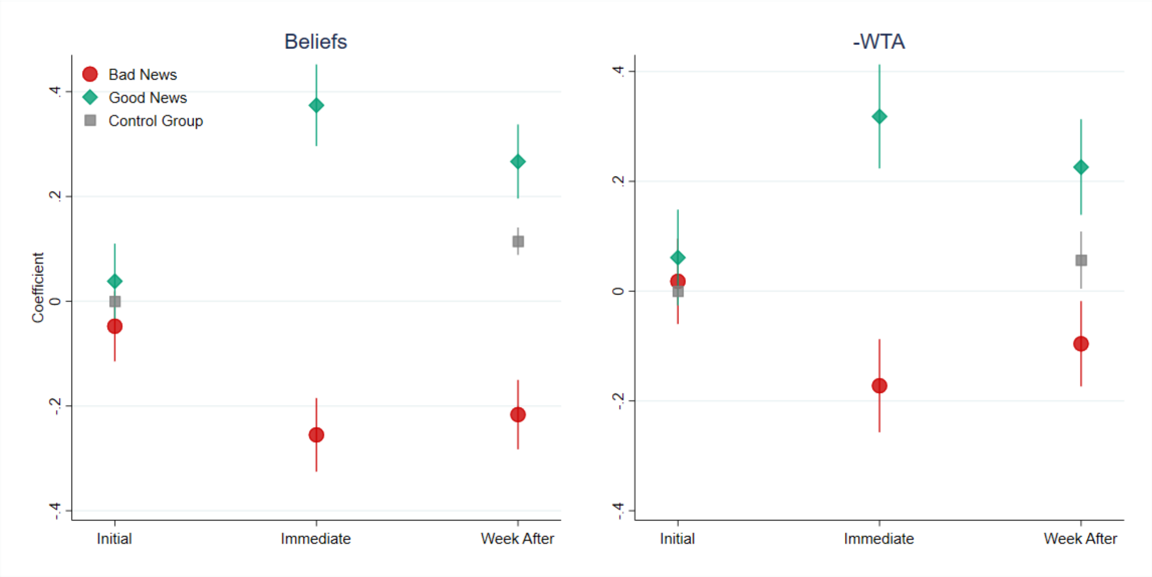


Figure 3.2: Price Lists for Round 2 Payments, Math



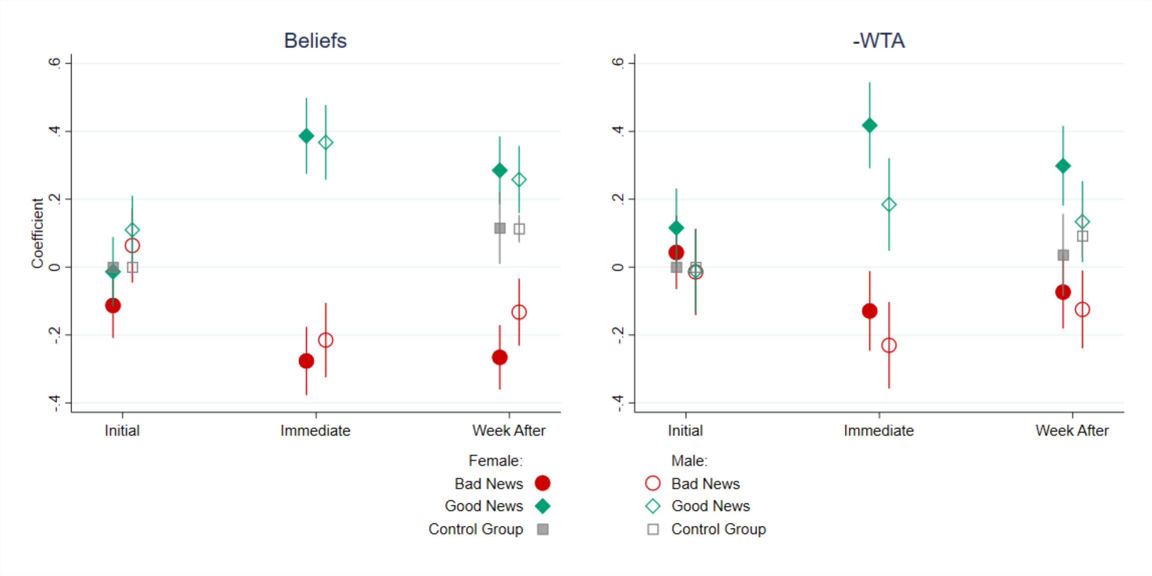
First Option	**Second Option**
Earn $4 per problem solved correctly on the **Math** test in **Round 2** if in top-4; $0 otherwise.	$1 per problem solved correctly on the **Verbal** test in **Round 2**.
Earn $3.50 per problem solved correctly on the **Math** test in **Round 2** if in top-4; $0 otherwise.	$1 per problem solved correctly on the **Verbal** test in **Round 2**.
Earn $3 per problem solved correctly on the **Math** test in **Round 2** if in top-4; $0 otherwise.	$1 per problem solved correctly on the **Verbal** test in **Round 2**.
Earn $2.50 per problem solved correctly on the **Math** test in **Round 2** if in top-4; $0 otherwise.	$1 per problem solved correctly on the **Verbal** test in **Round 2**.
Earn $2 per problem solved correctly on the **Math** test in **Round 2** test if in top-4; $0 otherwise.	$1 per problem solved correctly on the **Verbal** test in **Round 2**.
Earn $1.50 per problem solved correctly on the **Math** test in **Round 2** if in top-4; $0 otherwise.	$1 per problem solved correctly on the **Verbal** test in **Round 2**.

Figure 3.3: Levels over Time



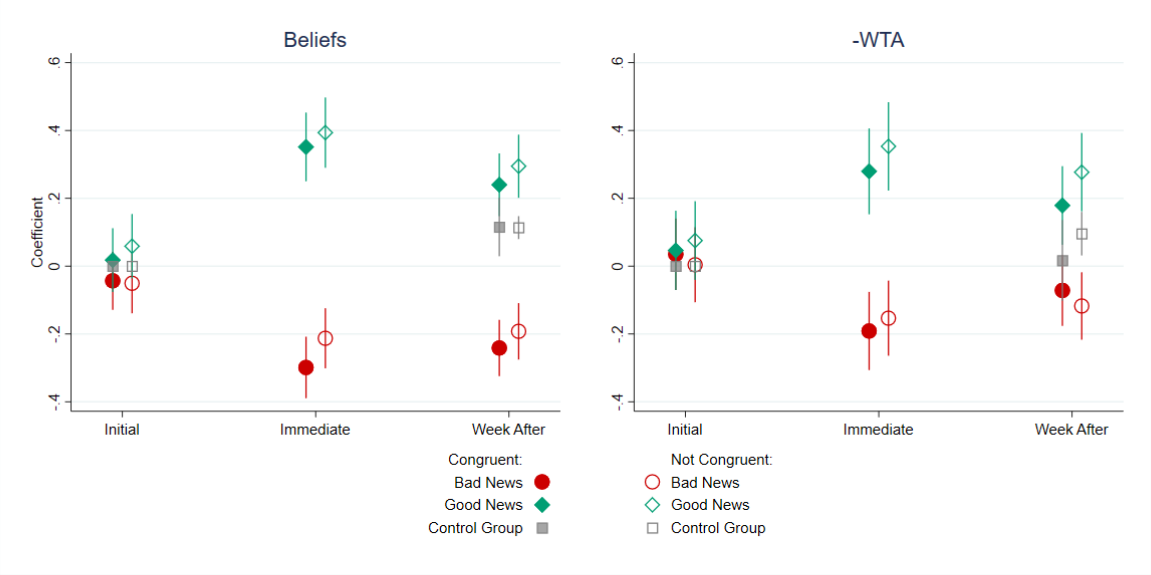
Note: Markers represent the coefficient on good and bad news, and control group at different stages from a regression that pools all stages for all the participants. The outcome, indicated at the top of each panel, is regressed on indicator variables for initial good news, immediate good news, immediate bad news, week after good news, week after bad news and week after control group (i.e. the omitted category is initial control group), $\boldsymbol{Y}$: relative and absolute performance controls, an indicator variable for the difficulty level of the first round tests and all the controls in $\boldsymbol{X}$: gender, family income, indicators for each parent attending college, a nonwhite indicator, ACT scores, high school rank, indicator for attending high school in the U.S., honors student indicator, school year (freshman, sophomore, junior or senior), a measure of risk aversion, an indicator for taking the math quiz first, domain indicator, and an immediate group indicator. Errors clustered at individual level. Beliefs is a composite variable that aggregates the three different measures of beliefs elicited, it is standardized with mean zero and standard deviation one at every stage. -WTA is the negative of the WTA standardized per stage. The spikes represent 90% confidence intervals.

Figure 3.4: Levels over Time by Gender



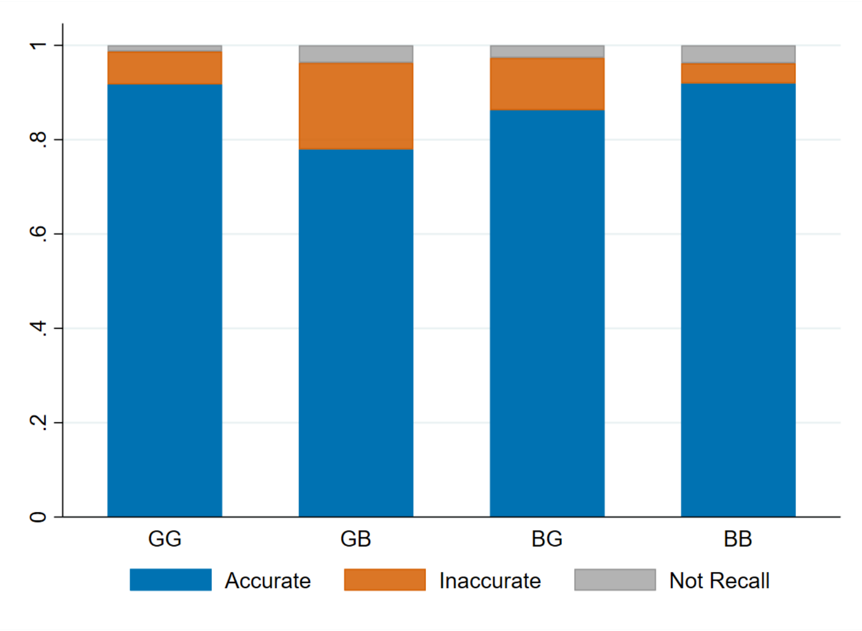
Note: Markers represent the coefficient on good and bad news, and control group at different stages for each gender from a regression that pools all stages for all the participants. The outcome, indicated at the top of each panel, is regressed on indicator variables for initial good news male and female, immediate good news male and female, immediate bad news male and female, week after good news male and female, week after bad news male and female, initial control female and week after control group male and female (i.e. the omitted category is initial control male), ***Y***: relative and absolute performance controls, an indicator variable for the difficulty level of the first round tests and all the controls in ***X***: family income, indicators for each parent attending college, a nonwhite indicator, ACT scores, high school rank, indicator for attending high school in the U.S., honors student indicator, school year (freshman, sophomore, junior or senior), a measure of risk aversion, an indicator for taking the math quiz first, domain indicator, and an immediate group indicator. The initial control female coefficient is normalized to zero, and all the female coefficients adjusted accordingly to be relative to the initial control female group. Errors clustered at individual level. Beliefs is a composite variable that aggregates the three different measures of beliefs elicited, it is standardized with mean zero and standard deviation one at every stage. -WTA is the negative of the WTA standardized per stage. The spikes represent 90% confidence intervals.

Figure 3.5: Levels over Time by Domain Congruence



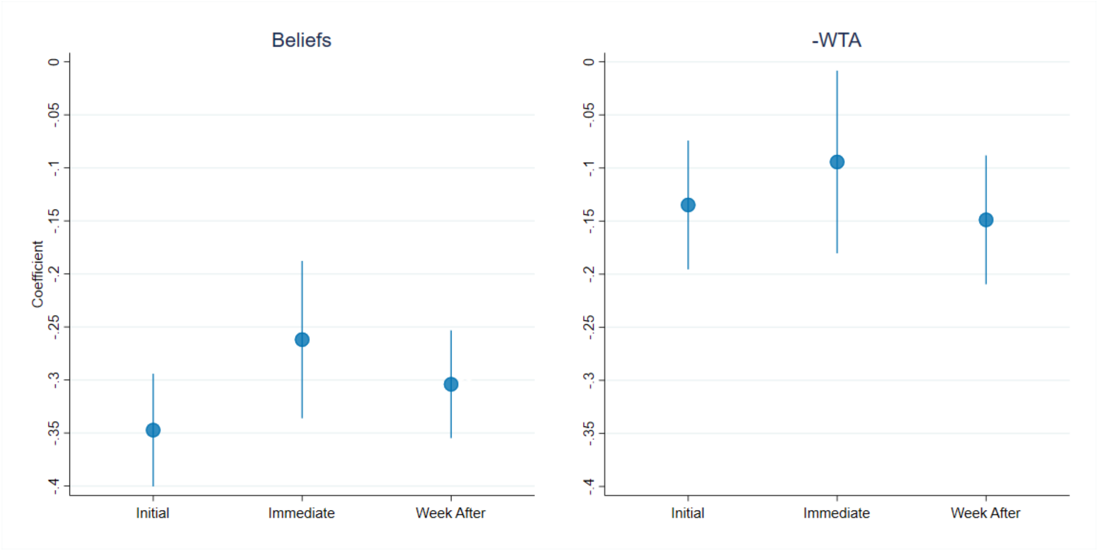
Note: Markers represent the coefficient on good and bad news, and control group at different stages for congruent and incongruent domains from a regression that pools all stages for all the participants. The outcome, indicated at the top of each panel, is regressed on indicator variables for initial good news congruent and incongruent, immediate good news congruent and incongruent, immediate bad news congruent and incongruent, week after good news congruent and incongruent, week after bad news congruent and incongruent, initial control congruent and week after control group congruent and incongruent (i.e. the omitted category is initial control not congruent), $\boldsymbol{Y}$: relative and absolute performance controls, an indicator variable for the difficulty level of the first round tests and all the controls in $\boldsymbol{X}$: gender, family income, indicators for each parent attending college, a nonwhite indicator, ACT scores, high school rank, indicator for attending high school in the U.S., honors student indicator, school year (freshman, sophomore, junior or senior), a measure of risk aversion, an indicator for taking the math quiz first, and an immediate group indicator. The initial control congruent coefficient is normalized to zero, and all the congruent coefficients adjusted accordingly to be relative to the initial control congruent category.Errors clustered at individual level. Beliefs is a composite variable that aggregates the three different measures of beliefs elicited, it is standardized with mean zero and standard deviation one at every stage. -WTA is the negative of the WTA standardized per stage. The spikes represent 90% confidence intervals.

Figure 3.6: Feedback Recall by Type of News



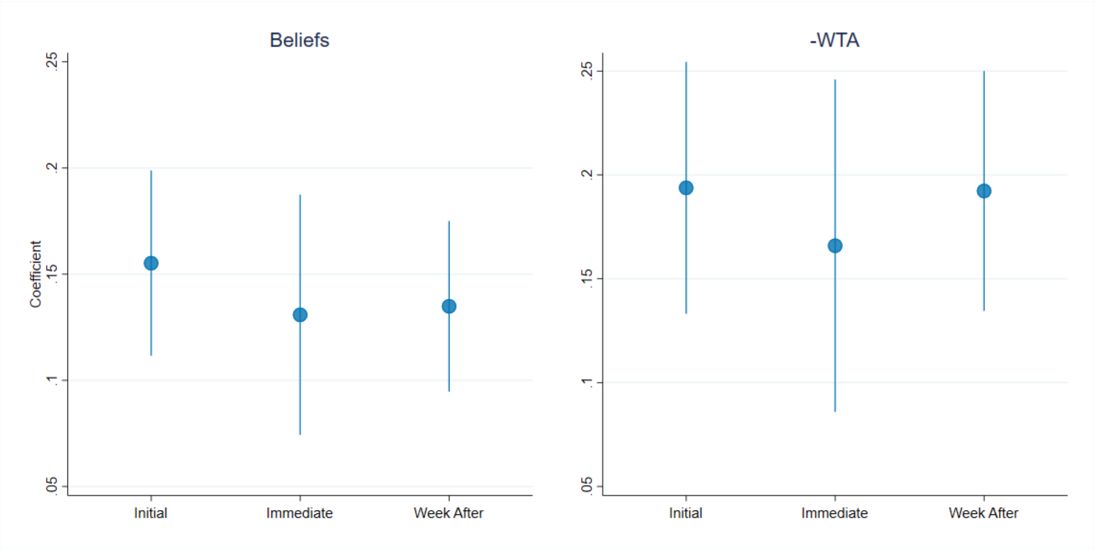
Note: GG: good news in both domains. BB: bad news in both domains. GB: good news in *D*, bad news in -*D*. BG: bad news in *D*, good news in -*D*. *D* denotes an observation associated with a given domain, and -*D* denotes the other domain. Standard errors reported in parentheses.

Figure 3.7: Male-Female Gap over Time



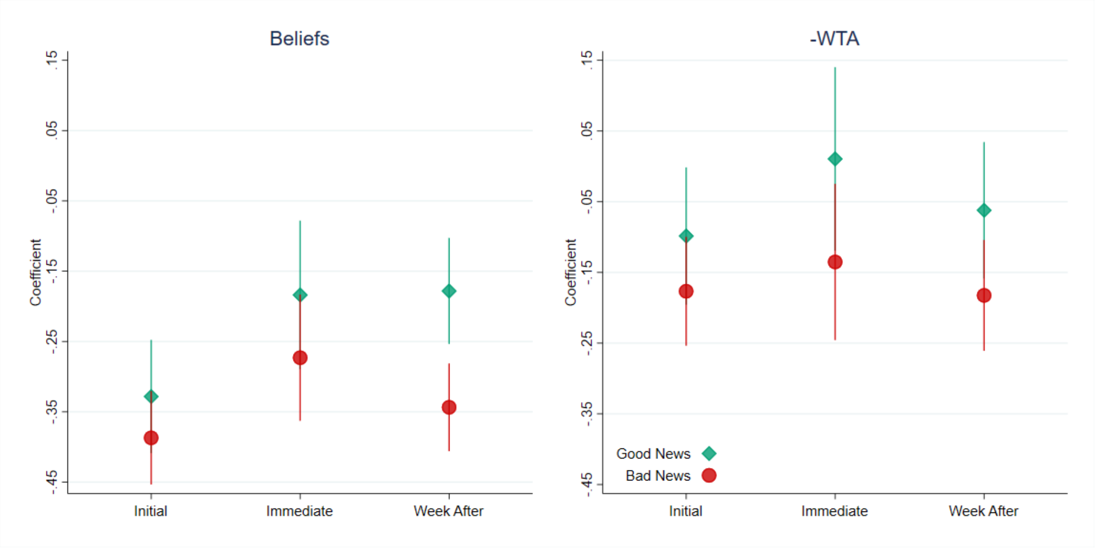
Note: Markers represent the coefficient on the female indicator interacted with initial, immediate and week after indicators (for females not in the control group) from a regression that pools all stages for all the participants. The outcome, indicated at the top of each panel, is regressed on indicators for immediate treated, week after treated, initial control, week after control (i.e. omitted category initial treated) and female interacted with those: initial treated female, immediate treated female, week after treated female, initial control female, week after control female, $\boldsymbol{Y}$: relative and absolute performance controls, an indicator variable for the difficulty level of the first round tests and all the controls in $\boldsymbol{X}$: family income, indicators for each parent attending college, a nonwhite indicator, ACT scores, high school rank, indicator for attending high school in the U.S., honors student indicator, school year (freshman, sophomore, junior or senior), a measure of risk aversion, an indicator for taking the math quiz first, an indicator for being in the Immediate group, and an domain congruence indicator. Errors clustered at individual level. Beliefs is a composite variable that aggregates the three different measures of beliefs elicited, it is standardized with mean zero and standard deviation one at every stage. -WTA is the negative of the WTA standardized per stage. The spikes represent 90% confidence intervals.

Figure 3.8: Stereotype Differences over Time



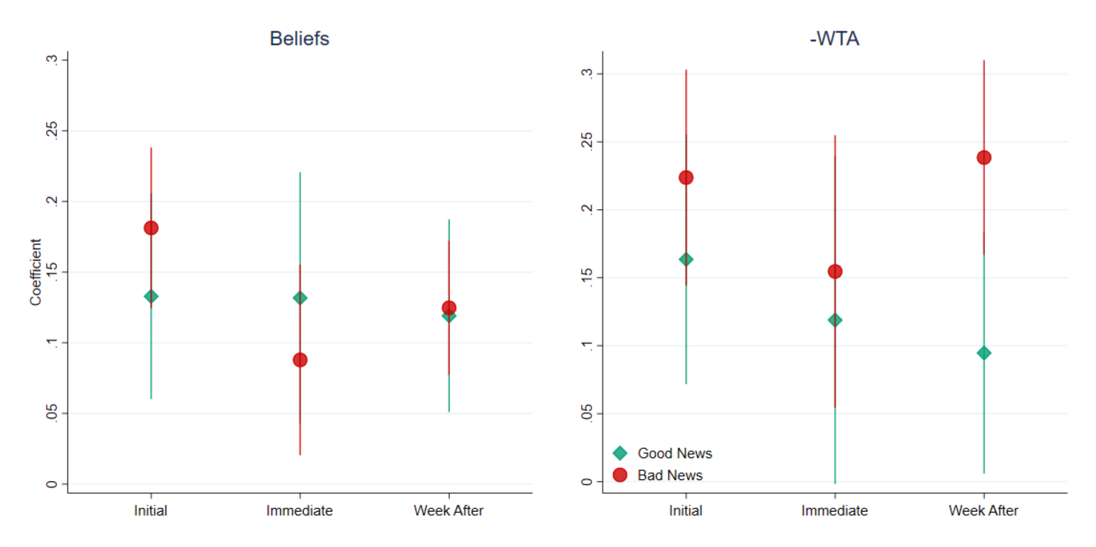
Note: Markers represent the coefficient on the domain congruence indicator interacted with initial, immediate and week after indicators (for participants not in the control group) from a regression that pools all stages for all the participants. The outcome, indicated at the top of each panel, is regressed on indicators for immediate treated, week after treated, initial control, week after control (i.e. omitted category initial treated) and domain congruence interacted with those: initial treated congruent, immediate treated congruent, week after treated congruent,initial control congruent, week after control congruent, $\boldsymbol{Y}$: relative and absolute performance controls, an indicator variable for the difficulty level of the first round tests and all the controls in $\boldsymbol{X}$: gender, family income, indicators for each parent attending college, a nonwhite indicator, ACT scores, high school rank, indicator for attending high school in the U.S., honors student indicator, school year (freshman, sophomore, junior or senior), a measure of risk aversion, an indicator for taking the math quiz first, and an indicator for being in the Immediate group. Errors clustered at individual level. Beliefs is a composite variable that aggregates the three different measures of beliefs elicited, it is standardized with mean zero and standard deviation one at every stage. -WTA is the negative of the WTA standardized per stage. The spikes represent 90% confidence intervals.

Figure 3.9: Male-Female Gap over Time by Type of News



Note: Markers represent the coefficient on the female indicator interacted with good and bad news at every stage from a regression that pools all stages for all the participants. The outcome, indicated at the top of each panel, is regressed on indicators for immediate treated good news, immediate treated bad news, week after treated good news, week after treated bad news, initial control, week after control (i.e. omitted category initial treated good news) and female interacted with those: initial treated good news female, initial treated bad news female, immediate treated good news female, immediate treated bad news female, week after treated good news female, week after treated bad news female, initial control female, week after control female, $\boldsymbol{Y}$: relative and absolute performance controls, an indicator variable for the difficulty level of the first round tests and all the controls in $\boldsymbol{X}$: family income, indicators for each parent attending college, a nonwhite indicator, ACT scores, high school rank, indicator for attending high school in the U.S., honors student indicator, school year (freshman, sophomore, junior or senior), a measure of risk aversion, an indicator for taking the math quiz first, an indicator for being in the Immediate group, and an domain congruence indicator. Errors clustered at individual level. Beliefs is a composite variable that aggregates the three different measures of beliefs elicited, it is standardized with mean zero and standard deviation one at every stage. -WTA is the negative of the WTA standardized per stage. The spikes represent 90% confidence intervals.

Figure 3.10: Stereotype Differences over Time by Type of News



Note: Markers represent the coefficient on the congruence indicator interacted with good and bad news at every stage from a regression that pools all stages for all the participants. The outcome, indicated at the top of each panel, is regressed on indicators for immediate treated good news, immediate treated bad news, week after treated good news, week after treated bad news, initial control, week after control (i.e. omitted category initial treated good news) and congruent interacted with those: initial treated good news congruent, initial treated bad news congruent, immediate treated good news congruent, immediate treated bad news congruent, week after treated good news congruent, week after treated bad news congruent, initial control congruent, week after control congruent, $\boldsymbol{Y}$: relative and absolute performance controls, an indicator variable for the difficulty level of the first round tests and all the controls in $\boldsymbol{X}$: family income, indicators for each parent attending college, a nonwhite indicator, ACT scores, high school rank, indicator for attending high school in the U.S., honors student indicator, school year (freshman, sophomore, junior or senior), a measure of risk aversion, an indicator for taking the math quiz first, an indicator for being in the Immediate group, and an domain congruence indicator. Errors clustered at individual level. Beliefs is a composite variable that aggregates the three different measures of beliefs elicited, it is standardized with mean zero and standard deviation one at every stage. -WTA is the negative of the WTA standardized per stage. The spikes represent 90% confidence intervals.

Table 3.1: Summary Statistics

	Female	Male	P-value
	(1)	(2)	(3)
Panel A			
% White	66.64	68.42	0.428
% Asian	21.02	23.55	0.205
% Black	3.93	2.63	0.136
% Hispanic	17.09	13.16	0.023
% First Generation	27.06	20.36	0.001
Family income[a]	103.97	118.34	0.000
Risk aversion[b]	3.66	3.31	0.000
GPA	3.69	3.64	0.013
ACT	22.75	25.31	0.000
% Honors	58.04	60.94	0.219
% Freshman	28.43	23.41	0.018
% Sophomore	23.49	29.50	0.004
% Junior	24.68	25.35	0.748
% Senior	23.40	21.75	0.410
Panel B: Experiment			
Score Math R1	5.74	6.49	0.000
Score Verbal R1	6.94	7.17	0.070
Score Math R2	4.18	4.77	0.000
Score Verbal R2	4.87	5.11	0.011
% Top-4 Math R1	17.92	31.16	0.000
% Top-4 Verbal R1	24.13	27.29	0.131
% Top-4 Math R2	15.90	23.41	0.000
% Top-4 Verbal R2	35.19	37.53	0.309
% Hard version	48.35	49.72	0.568
Beliefs Before Feedback			
Math guessed score	6.36	7.73	0.000
Verbal guessed score	7.18	7.62	0.000
Math confidence	63.87	66.22	0.030
Verbal confidence	63.39	61.25	0.039
Math guessed rank	5.26	6.78	0.000
Verbal guessed rank	5.93	6.57	0.000
Top-4 Math	32.58	53.31	0.000
Top-4 Verbal	38.79	49.57	0.000
Choices Before Feedback			
WTA Math[c]	337.51	285.14	0.000
WTA Verbal[c]	297.09	296.51	0.912
% Chose Math	39.67	56.23	0.000
N	1,094	722	

Note: Column (3) reports the p-value of a difference in means test across genders. Mean is reported for continuous variables. % Top-4 Math(Verbal) R1(R2) is the percentage of participants that scored in the top-4 when compared to the reference group in the math (verbal) quiz in Round 1 (Round 2). % Hard is the percentage of participants that got the hard version of the quizzes in Round 1. In the subpanel *Beliefs Before Feedback*, guessed scores range is 0-12, for confidence is 0-100 and for rank variables 1-10 where 10 is the best position. Top-4 Math (Verbal) is the mean belief (0-100) of being in the top-4 in the math (verbal) quiz. % Chose Math is the percentage of participantes to that prefer to be paid by their performance in math rather than verbal in Round 2.

[a] Family income in thousands of dollars.

[b] The higher the more risk averse (1-7).

[c] WTA in cents.

Table 3.2: Direction of Change in Rank Beliefs by Feedback Type

	Immediate		Week After	
	Good News	Bad News	Good News	Bad News
	(1)	(2)	(3)	(4)
Adjusted Up	0.40	0.09	0.38	0.19
No Change	0.50	0.33	0.45	0.38
Adjusted Down	0.10	0.58	0.18	0.43

Note: The table reports the proportion of participants that updated up, down or did not change their rank belief guess after receiving feedback. The immediate change (columns 1,2) is the calculated only for participants in the immediate group as the immediate measures minus the initial. The week after change (columns 3,4) is calculated as week after measure minus initial measure. The shaded cells represent proportion of participants for which the direction of the change in beliefs is what we would expect given the type of feedback.

Table 3.3: Feedback Recall

	(1)	(2)	(3)	(4)	(5)
Female (F)	0.056***	0.058***	0.073***	0.073***	0.056***
	(0.016)	(0.015)	(0.018)	(0.018)	(0.016)
Congruent	0.010	0.009	0.010	-0.002	0.009
	(0.009)	(0.009)	(0.009)	(0.011)	(0.009)
Good News$_D$	-0.093***	-0.170***	-0.068***	-0.085***	-0.098***
	(0.019)	(0.023)	(0.025)	(0.029)	(0.023)
Good News$_{-D}$		-0.088***			
		(0.020)			
Good News$_D$*Good News$_{-D}$		0.205***			
		(0.032)			
Good News$_D$*F			-0.043	-0.043	
			(0.027)	(0.027)	
Good News$_D$*Congruent				0.032	
				(0.027)	
Good News$_D$*Immediate group					0.010
					(0.026)
Immediate group	-0.009	-0.005	-0.009	-0.009	-0.013
	(0.014)	(0.014)	(0.014)	(0.014)	(0.016)
Mean	0.88	0.88	0.88	0.88	0.88
R2	0.04	0.06	0.04	0.04	0.04
Clusters	1,453	1,453	1,453	1,453	1,453
Obs.	2,906	2,906	2,906	2,906	2,906

Note: Outcome variable equals 1 if feedback is accurately recalled, 0 otherwise. All specifications control for ***Y***: relative and absolute performance controls, and an indicator variable for the difficulty level of the first round tests; time spent during Session 1, and ***X***: family income, indicators for each parent attending college, a nonwhite indicator, ACT scores, high school rank, indicator for attending high school in the U.S., honors student indicator, school year (freshman, sophomore, junior or senior), a measure of risk aversion, an indicator for taking the math quiz first and an indicator for being in the Immediate group. *D* denotes an observation associated with a given domain, and -*D* denotes the other domain. Standard errors reported in parentheses. Errors clustered at individual level. *Significant at 10%, **5%, ***1%.

Table 3.4: Effect of Priors

	Initial	Immediately After		Week After	
	(1)	(2)	(3)	(4)	(5)
Panel A: Beliefs					
Bad News$_D$	0.004	-0.660***	-0.645***	-0.449***	-0.453***
	(0.062)	(0.081)	(0.045)	(0.057)	(0.034)
Bad News$_D$*Female	-0.383***	-0.289***	-0.081**	-0.347***	-0.068***
	(0.041)	(0.056)	(0.034)	(0.038)	(0.024)
Good News$_D$*Female	-0.326***	-0.205***	0.021	-0.179***	0.059**
	(0.049)	(0.065)	(0.030)	(0.046)	(0.026)
Bad News$_D$*Congruent	0.173***	0.091**	-0.005	0.130***	0.004
	(0.035)	(0.042)	(0.022)	(0.029)	(0.017)
Good News$_D$*Congruent	0.127***	0.137**	0.008	0.122***	0.030
	(0.045)	(0.054)	(0.026)	(0.041)	(0.023)
Prior Beliefs			✓		✓
Mean	0.00	-0.00	-0.00	0.00	0.00
R2	0.42	0.55	0.87	0.52	0.83
Clusters	1,453	689	689	1,453	1,453
Obs.	2,906	1,378	1,378	2,906	2,906
Panel B: - WTA					
Bad News$_D$	0.002	-0.560***	-0.441***	-0.388***	-0.379***
	(0.079)	(0.109)	(0.071)	(0.076)	(0.060)
Bad News$_D$*Female	-0.171***	-0.159**	-0.029	-0.186***	-0.084**
	(0.047)	(0.068)	(0.047)	(0.048)	(0.041)
Good News$_D$*Female	-0.100*	-0.021	0.058	-0.070	-0.017
	(0.060)	(0.079)	(0.048)	(0.059)	(0.043)
Bad News$_D$*Congruent	0.224***	0.154**	0.014	0.235***	0.101***
	(0.049)	(0.062)	(0.038)	(0.044)	(0.033)
Good News$_D$*Congruent	0.157***	0.117	0.026	0.096*	0.003
	(0.056)	(0.074)	(0.045)	(0.054)	(0.037)
Prior WTA			✓		✓
Mean	-0.00	0.00	0.00	0.00	0.00
R2	0.16	0.27	0.70	0.22	0.53
Clusters	1,361	653	653	1,361	1,361
Obs.	2,603	1,262	1,262	2,627	2,627

Note: Outcome variable in Panel A is a composite variable that aggregates the three different measures of beliefs elicited, it is standardized with mean zero and standard deviation one at every stage. Outcome variable in Panel B is the negative of the WTA standarized per stage. Outcomes are regressed on an indicator for bad news, and interactions of good and bad news with female and congruent (i.e. omitted category is males who receive good news in gender-incongruent domains). All specifications control for $\boldsymbol{Y}$: relative and absolute performance controls, and an indicator variable for the difficulty level of the first round tests; and $\boldsymbol{X}$: family income, indicators for each parent attending college, a nonwhite indicator, ACT scores, high school rank, indicator for attending high school in the U.S., honors student indicator, school year (freshman, sophomore, junior or senior), a measure of risk aversion, an indicator for taking the math quiz first and an indicator for being in Immediate group. D denotes an observation associated with a given domain, and -D denotes the other domain. Errors clustered at individual level. Standard errors reported in parentheses. *Significant at 10%, **5%, ***1%.

REFERENCES

Ahn, T., P. Arcidiacono, A. Hopson and J. Thomas, "Equilibrium grade inflation with implications for female interest in stem majors", Working Paper (2022).

Ahn, T., P. Arcidiacono, A. Hopson and J. R. Thomas, "Equilibrium grade inflation with implications for female interest in stem majors", NBER Working Paper 26556, National Bureau of Economic Research, URL `http://www.nber.org/papers/w26556` (2019).

Alam, A. and I. S. Tapia, *Mapping gender equality in STEM from school to work* (UNICEF, 2020), URL `https://www.unicef.org/globalinsight/media/1361/file`.

Alston, M., "The (perceived) cost of being female: An experimental investigation of strategic responses to discrimination", Working Paper (2019).

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

Altonji, J. G., E. Blom and C. Meghir, "Heterogeneity in human capital investments: High school curriculum, college major, and careers", Annual Review of Economics **4**, 185–223, URL `http://www.annualreviews.org` (2012).

Altonji, J. G., L. B. Kahn and J. D. Speer, "Trends in earnings differentials across college majors and the changing task composition of jobs", American Economic Review **104**, 387–393, URL `https://web-s-ebscohost-com.ezproxy1.lib.asu.edu/ehost/pdfviewer/pdfviewer?vid=0&sid=8c9fe715-3d49-40b5-9099-f14f0aa4e9bc%40redis` (2014).

Aucejo, E. and J. James, "The path to college education: The role of math and verbal skills", Journal of Political Economy **129**, URL `https://eric.ed.gov` (2021).

Averett, C. P., "Block scheduling in north carolina high schools.", `https://files.eric.ed.gov/fulltext/ED406734.pdf` (1994).

Azmat, G. and N. Iriberri, "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, URL `https://EconPapers.repec.org/RePEc:eee:pubeco:v:94:y:2010:i:7-8:p:435-452` (2010).

Barber, B. M. and T. Odean, "Boys will be boys: Gender, overconfidence, and common stock investment", Quarterly Journal of Economics **116**, 261–292 (2001).

Barron, K., "Belief updating: does the 'good-news, bad-news' asymmetry extend to purely financial domains?", Experimental Economics **24** (2021).

Benjamin, D. J., “Chapter 2 - errors in probabilistic reasoning and judgment biases”, in “Handbook of Behavioral Economics - Foundations and Applications 2”, edited by B. D. Bernheim, S. DellaVigna and D. Laibson, vol. 2 of *Handbook of Behavioral Economics: Applications and Foundations 1*, pp. 69–186 (North-Holland, 2019), URL https://www.sciencedirect.com/science/article/pii/S2352239918300228.

Berlin, N. and M.-P. Dargnies, “Gender differences in reactions to feedback and willingness to compete”, Journal of Economic Behavior & Organization **130**, 320–336, URL http://www.sciencedirect.com/science/article/pii/S0167268116301615 (2016).

Beyer, S., “Gender differences in the accuracy of self-evaluations of performance.”, Journal of Personality and Social Psychology **59**, 5, 960–970 (1990).

Beyer, S., “Gender differences in self-perception and negative recall biases”, Sex Roles **38**, 103–133 (1998).

Beyer, S. and E. M. Bowden, “Gender differences in seff-perceptions: Convergent evidence from three measures of accuracy and bias”, Personality and Social Psychology Bulletin **23**, 2, 157–172, URL https://doi.org/10.1177/0146167297232005 (1997).

Blass, A. A., S. Lach and C. F. Manski, “Using elicited choice probabilities to estimate random utility models: Preferences for electricity reliability”, International Economic Review **51**, 421–440 (2010).

Bobba, M. and V. Frisancho, “Self-perceptions about academic achievement: Evidence from mexico city”, Journal of Econometrics (2021).

Bordalo, P., K. Coffman, N. Gennaioli and A. Shleifer, “Beliefs about Gender”, American Economic Review **109**, 3, 739–773 (2019).

Bordalo, P., N. Gennaioli and A. Shleifer, “Memory, Attention, and Choice”, The quarterly journal of economics **135**, 3, 1399–1442 (2020).

Breda, T. and C. Napp, “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, URL https://www.pnas.org/content/116/31/15435 (2019).

Brown, R., H. Mansour, S. D. O’Connell and J. Reeves, “Gender differences in political career progression: Evidence from u.s. elections”, Discussion Paper 12569, IZA (2019).

Buser, T., M. Niederle and H. Oosterbeek, “Gender, competitiveness, and career choices”, The Quarterly Journal of Economics **129**, 3, 1409–1447, URL https://EconPapers.repec.org/RePEc:oup:qjecon:v:129:y:2014:i:3:p:1409-1447 (2014).

Buser, T. and H. Yuan, “Do women give up competing more easily? evidence from the lab and the dutch math olympiad”, American Economic Journal: Applied Economics **11**, 3, 225–52, URL `https://www.aeaweb.org/articles?id=10.1257/app.20170160` (2019a).

Buser, T. and H. Yuan, “Do Women Give Up Competing More Easily? Evidence from the Lab and the Dutch Math Olympiad.”, American Economic Journal: Applied Economics **11**, 3, 225–52 (2019b).

Calonico, S., M. D. Cattaneo, M. H. Farrell and R. Titiunik, “Regression Discontinuity Designs Using Covariates”, The Review of Economics and Statistics **101**, 3, 442–451, URL `https://ideas.repec.org/a/tpr/restat/v101y2019i3p442-451.html` (2019).

Campbell, K. W. and C. Sedikides, “Self-Threat Magnifies the Self-Serving Bias: A Meta-Analytic Integration”, Review of General Psychology **3**, 1, 23–43 (1999).

Cason, T. N., W. A. Masters and R. M. Sheremeta, “Entry into winner-take-all and proportional-prize contests: An experimental study”, Journal of Public Economics **94**, 9, 604–611, URL `https://www.sciencedirect.com/science/article/pii/S0047272710000526` (2010).

Charness, G. and C. Dave, “Confirmation bias with motivated beliefs”, Games and Economic Behavior **104**, C, 1–23, URL `https://EconPapers.repec.org/RePEc:eee:gamebe:v:104:y:2017:i:c:p:1-23` (2017).

Chen, D. L., M. Schonger and C. Wickens, “otree—an open-source platform for laboratory, online, and field experiments”, Journal of Behavioral and Experimental Finance **9**, 88 – 97, URL `http://www.sciencedirect.com/science/article/pii/S2214635016000101` (2016).

Chew, S. H., W. Huang and X. Zhao, “Motivated False Memory”, Journal of Political Economy **128**, 10, 3913–3939, URL `https://ideas.repec.org/a/ucp/jpolec/doi10.1086-709971.html` (2020).

Coate, S. and G. C. Loury, “Will affirmative-action policies eliminate negative stereotypes?”, American Economic Review **83**, 1220–1240 (1993).

Coffman, K., M. Collis and L. Kulkarni, “Stereotypes and belief updating.”, Working Paper No. 19-068, Harvard Business School (2019).

Coffman, K., M. Collis and L. Kulkarni, “Stereotypes and belief updating”, Working paper (2021).

Coffman, K. B., “Evidence on Self-Stereotyping and the Contribution of Ideas”, The Quarterly Journal of Economics **129**, 4, 1625–1660, URL `https://ideas.repec.org/a/oup/qjecon/v129y2014i4p1625-1660.html` (2014).

Cortés, P., J. Pan, L. Pilossoph and B. Zafar, “Gender differences in job search and the earnings gap: Evidence from business majors”, Working Paper 28820, National Bureau of Economic Research (2021).

Coutts, A., "Good news and bad news are still news: Experimental evidence on belief updating.", Experimental Economics **22**, 2, 369 – 395, URL `http://search.ebscohost.com.ezproxy1.lib.asu.edu/login.aspx?direct=true&db=eoh&AN=1770987&site=ehost-live&scope=site` (2019).

De Paola, M. and F. Gioia, "Risk aversion and field of study choice: The role of individual ability", Bulletin of Economic Research **64**, 307–3378, URL `http://www.istat.it/lavoro/` (2012).

Dee, T. S., W. Dobbie, B. A. Jacob and J. Rockoff, "The Causes and Consequences of Test Score Manipulation: Evidence from the New York Regents Examinations", American Economic Journal: Applied Economics **11**, 3, 382–423, URL `https://ideas.repec.org/a/aea/aejapp/v11y2019i3p382-423.html` (2019).

Delaney, J. M. and P. J. Devereux, "Understanding gender differences in stem: Evidence from college applications.", Economics of Education Review **72**, 219–238, URL `http://www.sciencedirect.com/science/article/pii/S0272775719301761` (2019).

Delavande, A. and C. F. Manski, "Using elicited choice probabilities in hypothetical elections to study decisions to vote", Electoral Studies **38**, 28–37 (2015).

Eggers, A. C., R. Freier, V. Grembi and T. Nannicini, "Regression Discontinuity Designs Based on Population Thresholds: Pitfalls and Solutions", American Journal of Political Science **62**, 210–229 (2018).

Eil, D. and J. M. Rao, "The good news-bad news effect: Asymmetric processing of objective information about yourself", American Economic Journal: Microeconomics **3**, 2, 114–38, URL `https://www.aeaweb.org/articles?id=10.1257/mic.3.2.114` (2011).

Ellis, J., B. K. Fosdick and C. Rasmussen, "Women 1.5 times more likely to leave stem pipeline after calculus compared to men: Lack of mathematical confidence a potential culprit", PLOS ONE **11**, 7, 1–14, URL `https://doi.org/10.1371/journal.pone.0157447` (2016).

Ellison, G. and A. Swanson, "Dynamics of the gender gap in high math achievement", Working Paper 24910, National Bureau of Economic Research (2018).

Enke, B., F. Schwerter and F. Zimmermann, "Associative memory and belief formation", Working Paper 26664, National Bureau of Economic Research (2020).

Ertac, S. and B. Szentes, "The Effect of Information on Gender Differences in Competitiveness: Experimental Evidence", Koç University-TUSIAD Economic Research Forum Working Papers 1104, Koc University-TUSIAD Economic Research Forum, URL `https://ideas.repec.org/p/koc/wpaper/1104.html` (2011).

Exley, C. and J. Kessler, "The Gender Gap in Self-Promotion", Quarterly Journal of Economics (2022).

Fang, C., E. Zhang and J. Zhang, “Do women give up competing more easily? Evidence from speedcubers”, Economics Letters **205** (2021).

Fang, H. and A. Moro, “Theories of statistical discrimination and affirmative action: A survey”, Handbook of Social Economics **1**, 133–200 (2011).

Folke, O. and J. Rickne, “Sexual harassment and gender inequality in the labor market”, The Quarterly Journal of Economics **137**, 1–50 (2022).

Foschi, M., L. Lai and K. Sigerson, “Gender and double standards in the assessment of job applicants”, Social Psychology Quarterly **57**, 326–339 (1994).

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

Frandsen, B. R., “Party Bias in Union Representation Elections: Testing for Manipulation in the Regression Discontinuity Design when the Running Variable is Discrete”, in “Regression Discontinuity Designs”, vol. 38 of *Advances in Econometrics*, pp. 281–315 (Emerald Publishing Ltd, 2017), URL `https://ideas.repec.org/h/eme/aecozz/s0731-905320170000038012.html`.

Funk, C. and K. Parker, *Women and Men in STEM Often at Odds Over Workplace Equity*, vol. 9 (Pew Research Center, 2018), URL `www.pewresearch.org`.

Fuster, A., G. Kaplan and B. Zafar, “What would you do with $500? spending responses to gains, losses, news, and loans”, Review of Economic Studies **88**, 1760–1795, URL `https://academic.oup.com/restud/article/88/4/1760/5962017` (2021).

Fuster, A. and B. Zafar, “Survey experiments on economic expectations”, Handbook of Economic Expectations pp. 107–130 (2023).

Galindo-Silva, H., N. H. Somé and G. Tchuente, “Does Obamacare Care? A Fuzzy Difference-in-Discontinuities Approach”, GLO Discussion Paper Series 666, Global Labor Organization (GLO), URL `https://ideas.repec.org/p/zbw/glodps/666.html` (2020).

Ganley, C. M. and S. T. Lubienski, “Mathematics confidence, interest, and performance: Examining gender patterns and reciprocal relations”, Learning and Individual Differences **47**, 182–193, URL `http://www.sciencedirect.com/science/article/pii/S1041608016300024` (2016).

Gemici, A. and M. Wiswall, “Evolution of gender differences in post-secondary human capital investments: College majors”, International Economic Review **55** (2014).

Gentrup, S. and C. Rjosk, “Pygmalion and the gender gap: do teacher expectations contribute to differences in achievement between boys and girls at the beginning of schooling?”, Educational Research and Evaluation **24**, 3-5, 295–323 (2018).

Gill, D. and V. Prowse, “Gender differences and dynamics in competition: The role of luck”, Quantitative Economics **5**, 2, 351–376 (2014).

Golan, L. and C. Sanders, “Racial gaps, occupational matching, and skill uncertainty”, Federal Reserve Bank of St. Louis Review **101**, 135–153, URL `https://doi.org/10.20955/r.101.135` (2019).

Goldin, C., “Gender and the undergraduate economics major: Notes on the undergraduate economics major at a highly selective liberal arts college”, (2015).

Goldin, C. and C. Rouse, “Orchestrating impartiality: The impact of ’blind’ auditions on female musicians”, American Economic Review **90**, 715–741, URL `https://web-p-ebscohost-com.ezproxy1.lib.asu.edu/ehost/pdfviewer/pdfviewer?vid=0&sid=7749a872-7979-439c-8628-3dd2dfa6f077%40redis` (2000).

Gotthard-Real, A., “Desirability and information processing: An experimental study”, Economics Letters **152**, C, 96–99, URL `https://EconPapers.repec.org/RePEc:eee:ecolet:v:152:y:2017:i:c:p:96-99` (2017).

Grant Thornton International Ltd, “Women in business 2021: A window of opportunity”, (2021).

Grembi, V., T. Nannicini and U. Troiano, “Do Fiscal Rules Matter?”, American Economic Journal: Applied Economics **8**, 3, 1–30, URL `https://ideas.repec.org/a/aea/aejapp/v8y2016i3p1-30.html` (2016).

Grossman, Z. and D. Owens, “An unlucky feeling: Overconfidence and noisy feedback”, Journal of Economic Behavior & Organization **84**, 2, 510–524, URL `https://EconPapers.repec.org/RePEc:eee:jeborg:v:84:y:2012:i:2:p:510-524` (2012).

Hammond, A., E. R. Matulevich, K. Beegle and S. K. Kumaraswamy, *The Equality Equation: Advancing the Participation of Women and Girls in STEM* (World Bank, 2020).

Heider, F., “The Psychology of Interpersonal Relations”, Psychology Press (1958).

Huffman, D., C. Raymond and J. Shvets, “Persistent overconfidence and biased memory:evidence from managers”, Working paper (2019).

Hunt, J., “Why do women leave science and engineering?”, Industrial and Labor Relations Review **69**, 199–226 (2016).

Imbens, G. and K. Kalyanaraman, “Optimal Bandwidth Choice for the Regression Discontinuity Estimator”, Review of Economic Studies **79**, 3, 933–959, URL `https://ideas.repec.org/a/oup/restud/v79y2012i3p933-959.html` (2012).

Joensen, J. S. and H. S. Nielsen, “Is there a causal effect of high school math on labor market outcomes?”, The Journal of Human Resources **44**, 1, 171–198, URL `http://www.jstor.org/stable/20648891` (2009).

Kaganovich, M., M. Taylor and R. Xiao, “Gender differences in persistence in a field of study”, CESifo Working Paper Series 9087, CESifo (2021).

Kang, L., Z. Lei, Y. Song and P. Zhang, “Gender differences in reactions to failure in high-stakes competition: evidence from the national college entrance exam retakes”, Working paper, SSRN (2021).

Kessel, D., J. Mollerstrom and R. van Veldhuizen, “Can simple advice eliminate the gender gap in willingness to compete?”, European Economic Review **138** (2021).

Kessler, J. B., C. Low and C. D. Sullivan, “Incentivized resume rating: Eliciting employer preferences without deception”, American Economic Review **109**, 3713–3744 (2019).

Kiefer, A. and M. Shih, “Gender differences in persistence and attributions in stereotype relevant contexts.”, Sex Roles **54**, 859–868 (2006).

Koşar, G., T. Ransom and W. V. D. Klaauw, “Understanding migration aversion using elicited counterfactual choice probabilities”, Journal of Econometrics **231**, 123–147, URL `https://doi.org/10.1016/j.jeconom.2020.07.056` (2022).

Kugler, A. D., C. H. Tinsley and O. Ukhaneva, “Choice of majors: are women really different from men?”, Economics of Education Review **81**, 102079, URL `https://www.sciencedirect.com/science/article/pii/S0272775721000029` (2021).

Lavy, V. and E. Sand, “On the origins of gender gaps in human capital: Short- and long-term consequences of teachers’ biases”, Journal of Public Economics **167**, 263–279, URL `https://www.sciencedirect.com/science/article/pii/S0047272718301750` (2018).

Lee, D. S. and D. Card, “Regression discontinuity inference with specification error”, Journal of Econometrics **142**, 2, 655–674, URL `https://EconPapers.repec.org/RePEc:eee:econom:v:142:y:2008:i:2:p:655-674` (2008).

Levine, P. B. and D. J. Zimmerman, “The benefit of additional high-school math and science classes for young men and women”, Journal of Business & Economic Statistics **13**, 2, 137–149, URL `http://www.jstor.org/stable/1392368` (1995).

Long, M. C., D. Conger and P. Iatarola, “Effects of high school course-taking on secondary and postsecondary success”, American Educational Research Journal **49**, 2, 285–322, URL `http://www.jstor.org/stable/41419458` (2012).

Lundeberg, M., P. W. Fox and J. Punćcohaŕ, “Highly confident, but wrong: Gender differences and similarities in confidence judgments.”, Journal of Educational Psychology **86**, 114–121 (1994).

Macartney, H., R. McMillan and U. Petronijevic, “Teacher performance and accountability incentives”, NBER Working Paper 24747, National Bureau of Economic Research, URL `http://www.nber.org/papers/w24747` (2018).

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

McCrary, J., “Manipulation of the running variable in the regression discontinuity design: A density test”, Journal of Econometrics **142**, 2, 698–714, URL http://www.sciencedirect.com/science/article/pii/S0304407607001133 (2008).

McKinney, A. P., K. D. Carlson, R. L. I. Mecham, N. D. D. Angelo and M. L. Connerley, “Recruiters’ use of gpa in initial screening decisions: Higher gpas don’t always make the cut”, Personnel Psychology **56**, 823–845 (2003).

McKinney, A. P. and A. Miles, “Gender differences in us performance measures for personnel selection”, Equal Opportunities International **28**, 121–134, URL https://www-emerald-com.ezproxy1.lib.asu.edu/insight/content/doi/10.1108/02610150910937880/full/pdf (2009).

Mezulis, A., L. Abramson, J. Hyde and B. Hankin, “Is there a universal positivity bias in attributions? A meta-analytic review of individual, developmental, and cultural differences in the self-serving attributional bias”, Psychological Bulletin **130**, 5, 711–747 (2004).

Miller, D. T. and M. Ross, “Self-serving biases in the attribution of causality: Fact or fiction?”, Psychological Bulletin **82**, 2, 213–225 (1975).

Moakler, M. W. J. and M. M. Kim, “College major choice in stem: Revisiting confidence and demographic factors”, The Career Development Quarterly **62**, 128–142, URL https://www.proquest.com/docview/1555300798?pq-origsite=gscholar&fromopenview=true (2014).

Mobius, M., M. Niederle, P. Niehaus and T. Rosenblat, “Managing self-confidence: Theory and experimental evidence”, Management Science (2021).

National Center for Education Statistics, “2015 survey questionnaires results: Students’ views of mathematics, reading, and science (nces 2018-155)”, https://www.nationsreportcard.gov/sq_students_views_2015/, URL https://www.nationsreportcard.gov/sq_students_views_2015/ (2018).

National Center for Education Statistics, “2015 survey questionnaires results: Classroom instructions for mathematics, reading, and science (nces 2018-147)”, https://www.nationsreportcard.gov/sq_classroom/#mathematics, URL https://www.nationsreportcard.gov/sq_students_views_2015/ (2019).

Niederle, M. and L. Vesterlund, “Do women shy away from competition? do men compete too much?*”, The Quarterly Journal of Economics **122**, 3, 1067–1101, URL http://dx.doi.org/10.1162/qjec.122.3.1067 (2007).

Nord, C., S. Roey, R. Perkins, M. Lyons, N. Lemanski, J. Brown and J. Schuknecht, “The nation’s report card: America’s high school graduates (nces 2011-462)”, (2011).

North Carolina Department of Public Instruction, “2013–14 school performance grades (a–f) for north carolina public schools”, https://www.dpi.nc.gov/media/5814/open (2015).

North Carolina Department of Public Instruction, "North carolina end-of-course assessment of nc math 1", http://www.ncpublicschools.org/docs/accountability/testing/achievelevels/math1achvlvl16.pdf (2016a).

North Carolina Department of Public Instruction, "The north carolina testing program technical report 2012–2015 english language arts/reading assessments (ela) end-of-grade 3–8 and end-of-course english ii", https://www.dpi.nc.gov/media/9619/open (2016b).

North Carolina Department of Public Instruction, "The north carolina testing program technical report 2012–2015 mathematics assessments end-of-grade 3–8 and end-of-course math i", https://files.nc.gov/dpi/documents/accountability/testing/technotes/mathtechreport1215.pdf (2016c).

North Carolina Department of Public Instruction, "End-of-grade (eog)", https://www.dpi.nc.gov/districts-schools/testing-and-school-accountability/state-tests/end-grade-eog/, URL https://www.dpi.nc.gov/districts-schools/testing-and-school-accountability/state-tests/end-grade-eog (2019a).

North Carolina Department of Public Instruction, "North carolina end-of-course tests", http://www.ncpublicschools.org/accountability/testing/eoc/ (2019b).

North Carolina Department of Public Instruction, "Course code manual revised march 2022", https://www.dpi.nc.gov/media/14282/open (2022).

Ost, B., "The role of peers and grades in determining major persistence in the sciences", Economics of Education Review **29**, 6, 923–934, URL https://www.sciencedirect.com/science/article/pii/S0272775710000762 (2010).

Owen, S., "College field specialization and beliefs about relative performance", Working paper (2020).

Papay, J. P., R. J. Murnane and J. B. Willett, "The impact of test score labels on human-capital investment decisions", Journal of Human Resources **51**, 357–388, URL http://jhr.uwpress.org/content/51/2/357http://jhr.uwpress.org/content/51/2/357.abstract (2016).

Patnaik, A., M. J. Wiswall and B. Zafar, "College majors", The Routledge Handbook of the Economics of Education **1**, URL https://www.nber.org/system/files/working_papers/w27645/w27645.pdf (2021).

Pereda, P., L. Matsunaga, M. D. M. Diaz, B. P. Borges, J. Mena-Chalco, F. Rocha, R. Narita and C. Brenck, "Are women less persistent? evidence from submissions to a nationwide meeting of economics", Working Paper 2020-19, FEA-USP (2020).

Pew Research Center, "The data on women leaders", URL https://www.pewresearch.org/social-trends/fact-sheet/the-data-on-women-leaders/ (2021).

Quadlin, N., “The mark of a woman’s record: Gender and academic performance in hiring”, American Sociological Review **83**, 331–360 (2018).

Quintero, E., *How are Job Applicants Disadvantaged by Gender Based Double Standards in a Natural Setting* (Ph.D. thesis, Cornell University, 2008), URL https://www.proquest.com/docview/304632466?parentSessionId=5HW%2B%2FUlzxDCAibFFyvAkS3JBkJ2GScsoDrSLjpJHcik%3D&pq-origsite=primo&accountid=4485https://ecommons.cornell.edu/handle/1813/11184.

Rahimi, E. and S. S. H. Nazari, “A detailed explanation and graphical representation of the blinder-oaxaca decomposition method with its application in health inequalities”, Emerging Themes in Epidemiology **18**, URL https://doi.org/10.1186/s12982-021-00100-9 (2021).

Rask, K. and J. Tiefenthaler, “The role of grade sensitivity in explaining the gender imbalance in undergraduate economics”, Economics of Education Review **27**, 6, 676–687, URL http://www.sciencedirect.com/science/article/pii/S0272775707001185 (2008).

Reuben, E., P. Sapienza and L. Zingales, “Taste for competition and the gender gap among young business professionals.”, Working paper (2019).

Reuben, E., M. Wiswall and B. Zafar, “Preferences and biases in educational choices and labour market expectations: Shrinking the black box of gender”, The Economic Journal **127**, 604, 2153–2186, URL https://onlinelibrary.wiley.com/doi/abs/10.1111/ecoj.12350 (2017).

Riegle-Crumb, C. and M. Humphries, “Exploring bias in math teachers’ perceptions of students’ ability by gender and race/ethnicity”, Gender & Society **26**, 2, 290–322, URL https://doi.org/10.1177/0891243211434614 (2012).

Roberts, M. R. and T. M. Whited, “Chapter 7 - endogeneity in empirical corporate finance1”, vol. 2 of *Handbook of the Economics of Finance*, pp. 493–572 (Elsevier, 2013), URL http://www.sciencedirect.com/science/article/pii/B9780444535948000070.

Roberts, T.-A. and S. Nolem-Hoeksema, “Sex differences in reactions to evaluative feedback”, Sex Roles **21**, 725–747 (1989).

Rose, H. and J. R. Betts, “The effect of high school courses on earnings”, The Review of Economics and Statistics **86**, 2, 497–513, URL https://doi.org/10.1162/003465304323031076 (2004).

Roth, C., I. Haaland and J. Wohlfart, “Designing information provision experiments”, Journal of Economic Literature, forthcoming (2021).

Russel, K. and P. Carter, *Discover Your IQ Potentail: Over 500 Tests of Your Mental Agility* (Arcturus, 2001).

Schwardmann, P. and J. Weele, “Deception and self-deception”, Nature Human Behaviour **3** (2019).

Shastry, G. K., O. Shurchkov and L. L. Xia, "Luck or skill: How women and men react to noisy feedback", Journal of Behavioral and Experimental Economics **88**, 101592, URL https://www.sciencedirect.com/science/article/pii/S2214804320301403 (2020).

Shettle, C., S. Roey, J. Mordica, R. Perkins, C. Nord, J. Teodorovic, J. Brown, M. Lyons, C. Averett and D. Kastberg, "The nation's report card: America's high school graduates (nces 2007-467)", (2007).

Steele, J., J. B. James and R. C. Barnett, "Learning in a man's world: Examining the perceptions of undergraduate women in male-dominated academic areas", Psychology of Women Quarterly **26**, 46–50 (2002).

Tan, B., "Grades as noisy signals", Pedagogy eJournal (2020).

Toft Hansen, A., U. Hvidman and H. H. Sievertsen, "Grades and employer learning", Journal of Labor Economics URL https://www-journals-uchicago-edu.ezproxy1.lib.asu.edu/doi/pdfplus/10.1086%2F724048 (2023).

Tungodden, J. and A. Willén, "When parents decide: Gender differences in competitiveness", CESifo Working Paper Series 9516, CESifo (2022).

Wasserman, M., "Gender Differences in Politician Persistence", Review of Economics and Statistics **Forthcoming** (2020).

Williams, J. C., K. W. Phillips and E. V. Hall, *Double Jeopardy? Gender Bias Against Women in Science* (Work Life Law, 2014), URL www.worklifelaw.org.

Wiswall, M. and B. Zafar, "Preference for the workplace, investment in human capital, and gender", Quarterly Journal of Economics **133**, 457–507 (2018).

Wozniak, D., W. T. Harbaugh and U. Mayr, "The menstrual cycle and performance feedback alter gender differences in competitive choices", Journal of Labor Economics **32**, 1, 161–198, URL http://www.jstor.org/stable/10.1086/673324 (2014).

Wright, C., *Choose Wisely: A Study of College Major Choice and Major Switching Behavior* (Ph.D. thesis, Pardee Rand Graduate School, 2018), URL www.rand.org/giving/contribute.

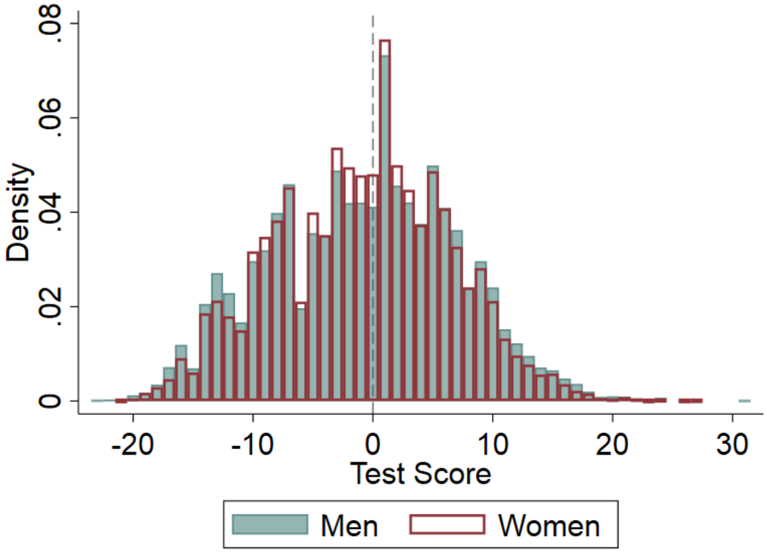
Zimmermann, F., "The dynamics of motivated beliefs", American Economic Review **110**, 2, 337–61, URL https://www.aeaweb.org/articles?id=10.1257/aer.20180728 (2020).

APPENDIX A

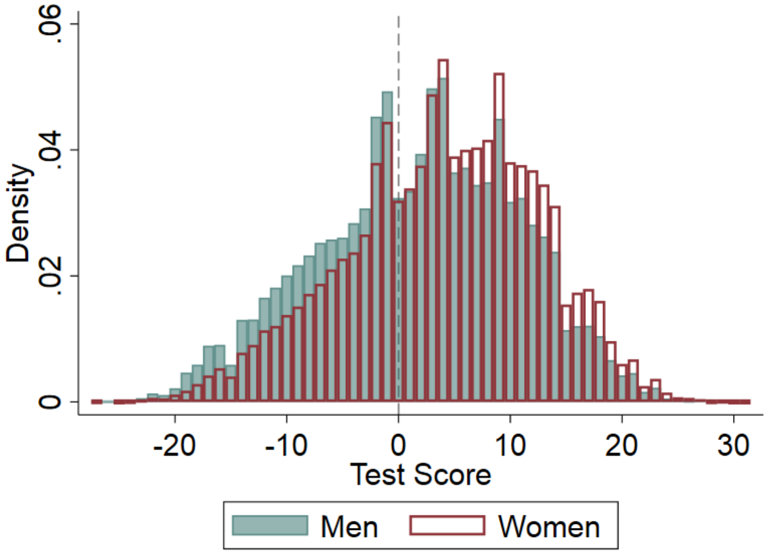
ADDITIONAL FIGURES AND TABLES, CHAPTER 1

Figure A.1: Test Scores Distributions

(a) Math 1.

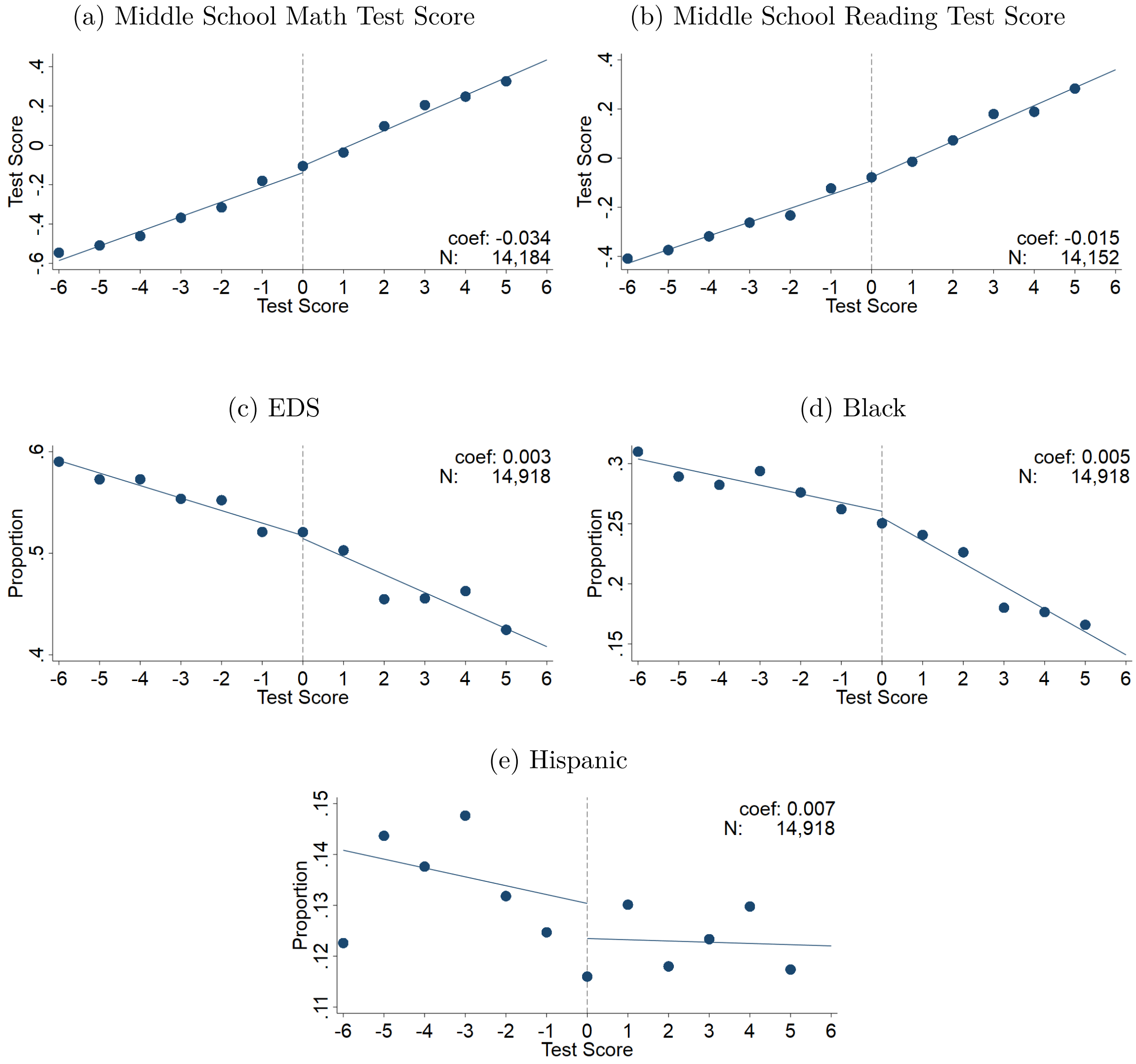


(b) English 2.



Note: Scores are normalized such that a score of 0 or more means proficiency.

Figure A.2: Continuity of Covariates for Math 1 Test



(a) Middle School Math Test Score

(b) Middle School Reading Test Score

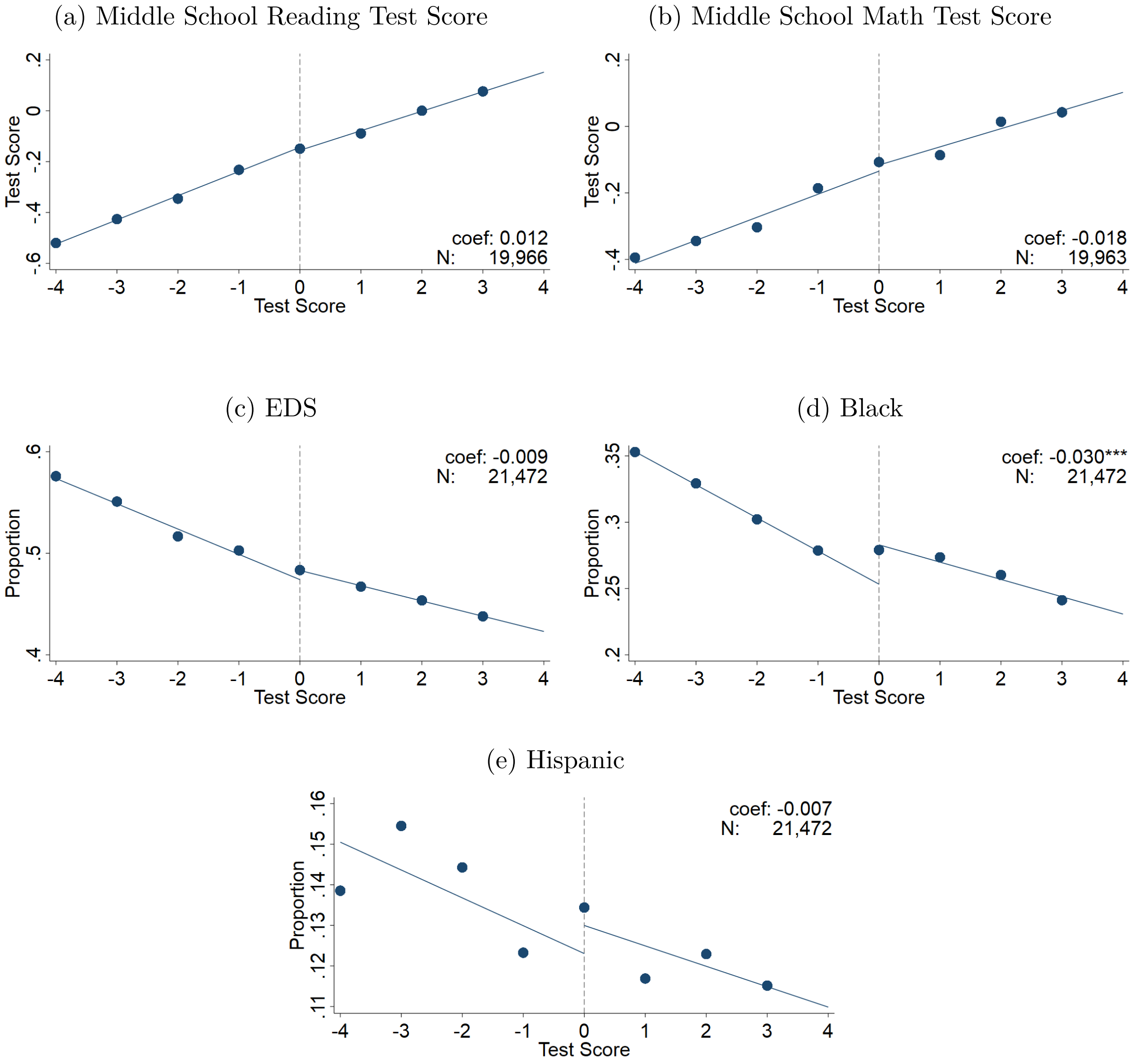
(c) EDS

(d) Black

(e) Hispanic

Note: Math test scores are normalized such that a score of 0 or more means proficiency. Lines are fitted values from regressing the corresponding covariate on an indicator for being non-proficient on the Math 1 test. Additionally all specifications include a first-degree polynomial of the running variable with different slopes above and below the proficiency threshold, within ±6 of the cutoff. Standard errors are clustered at test score level. Middle school test scores are standardized by year. The dots are averages within each 1 point bin. Minority: black or hispanic students. EDS: Economically Disadvantaged Student. Coef: estimated discontinuity, i.e. the coefficient for non-proficiency. *Significant at 10%, **5%, ***1%.

Figure A.3: Continuity of Covariates for English 2 Test



(a) Middle School Reading Test Score

(b) Middle School Math Test Score

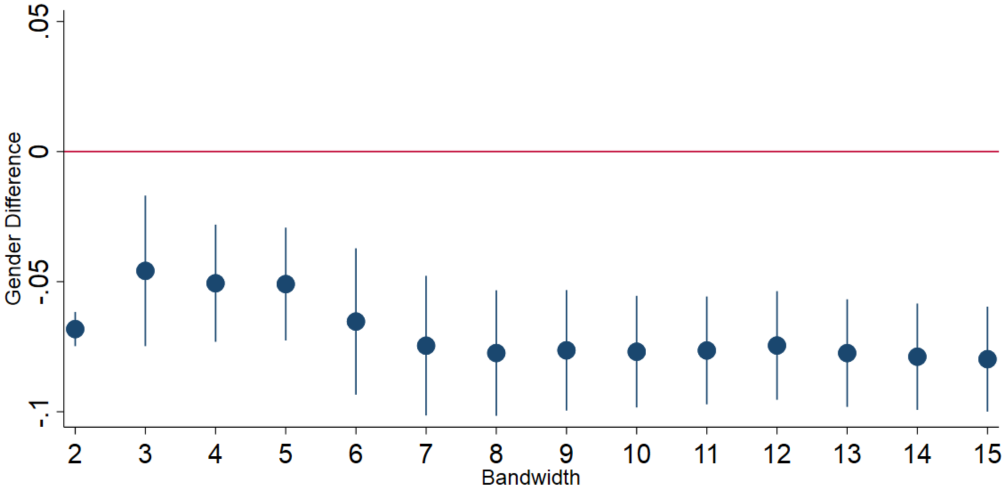
(c) EDS

(d) Black

(e) Hispanic

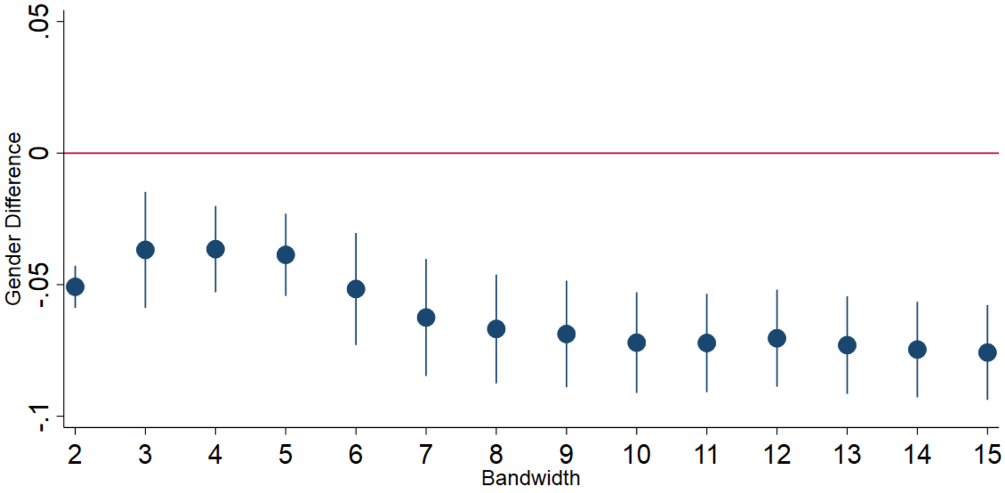
Note: English test scores are normalized such that a score of 0 or more means proficiency. Lines are fitted values from regressing the corresponding covariate on an indicator for being non-proficient on the English 2 test. Additionally all specifications include a first-degree polynomial of the running variable with different slopes above and below the proficiency threshold, within ±4 of the cutoff. Standard errors are clustered at test score level. Middle school test scores are standardized by year. The dots are averages within each 1 point bin. Minority: black or hispanic students. EDS: Economically Disadvantaged Student. Coef: estimated discontinuity, i.e. the coefficient for non-proficiency. *Significant at 10%, **5%, ***1%.

Figure A.4: Estimates of the Difference in the Discontinuity for Different Bandwidths, Math



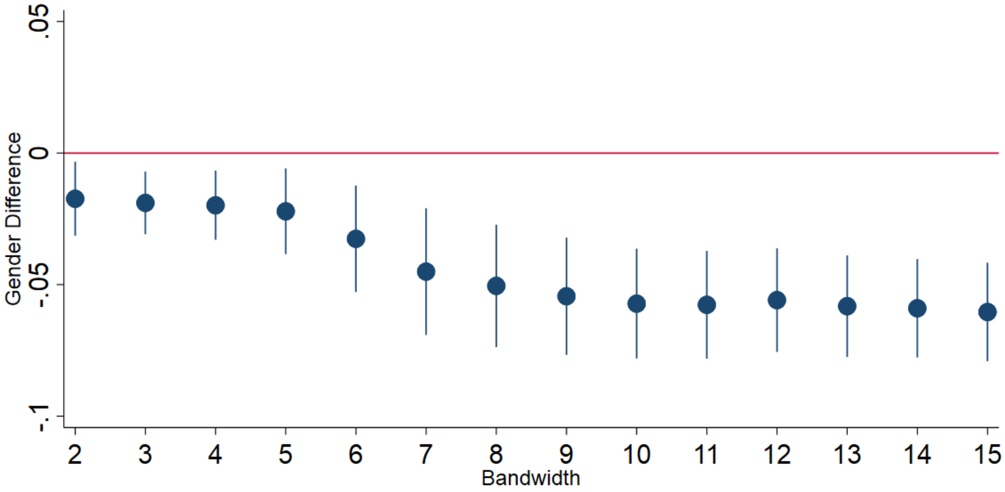
Note: Spikes represent 90% confidence intervals. Dots are the coefficients for being a non-proficient female from a regression of an indicator for taking honors Math 2 on indicators for female and being non-proficient on the Math 1 EOC test, the interaction of those two variables, a first-degree polynomial of the running variable with different slopes above and below the proficiency threshold. Standard errors are clustered at test score level. Optimal bandwidth is between 6 and 7.

Figure A.5: Estimates of the Difference in the Discontinuity for Different Bandwidths, Math



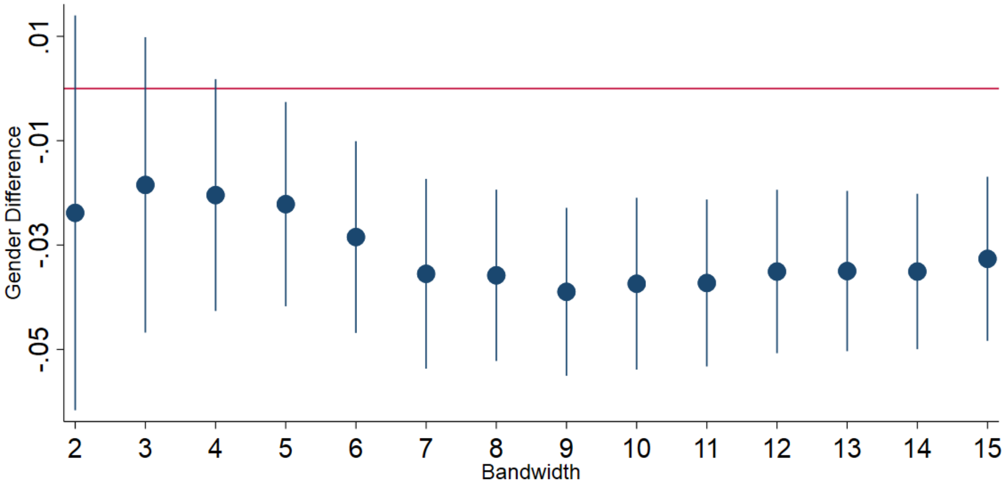
Note: Spikes represent 90% confidence intervals. Dots are the coefficients for being a non-proficient female from a regression of an indicator for taking honors Math 2 on indicators for female and being non-proficient on the Math 1 EOC test, the interaction of those two variables, a first-degree polynomial of the running variable with different slopes above and below the proficiency threshold, year and high school FE, indicators for EDS and race. Standard errors are clustered at test score level. Optimal bandwidth is between 6 and 7.

Figure A.6: Estimates of the Difference in the Discontinuity for Different Bandwidths, Math



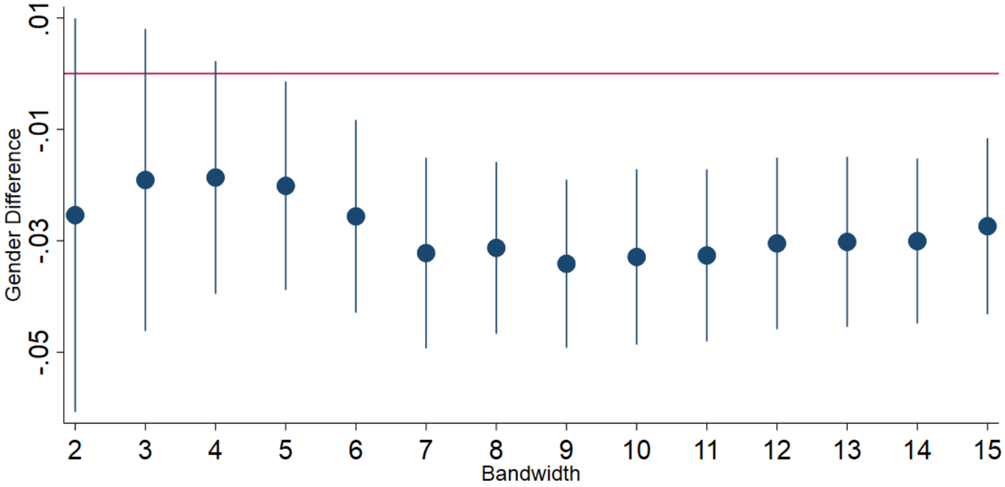
Note: Spikes represent 90% confidence intervals. Dots are the coefficients for being a non-proficient female from a regression of an indicator for taking honors Math 2 on indicators for female and being non-proficient on the Math 1 EOC test, the interaction of those two variables, a first-degree polynomial of the running variable with different slopes above and below the proficiency threshold, year and high school FE, indicators for EDS, race, and the difference between the standardized math and reading 8th grade test scores. Standard errors are clustered at test score level. Optimal bandwidth is between 6 and 7.

Figure A.7: Estimates of the Difference in the Discontinuity for Different Bandwidths, English



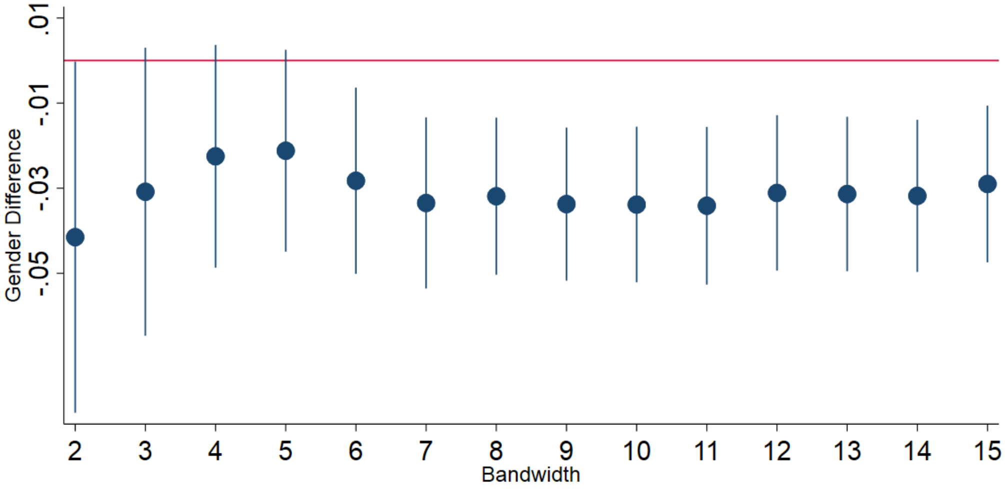
Note: Spikes represent 90% confidence intervals. Dots are the coefficients for being a non-proficient female from a regression of an indicator for taking advanced English 3 class on indicators for female, being non-proficient on the English 2 EOC test, the interaction of those two variables, a first-degree polynomial of the running variable with different slopes above and below the proficiency threshold. Standard errors are clustered at test score level. Optimal bandwidth is between 4 and 5.

Figure A.8: Estimates of the Difference in the Discontinuity for Different Bandwidths, English



Note: Spikes represent 90% confidence intervals. Dots are the coefficients for being a non-proficient female from a regression of an indicator for taking advanced English 3 class on indicators for female, being non-proficient on the English 2 EOC test, the interaction of those two variables, a first-degree polynomial of the running variable with different slopes above and below the proficiency threshold, high school and year FE, indicators for EDS and race. Standard errors are clustered at test score level. Optimal bandwidth is between 4 and 5.

Figure A.9: Estimates of the Difference in the Discontinuity for Different Bandwidths, English



Note: Spikes represent 90% confidence intervals. Dots are the coefficients for being a non-proficient female from a regression of an indicator for taking advanced English 3 class on indicators for female, being non-proficient on the English 2 EOC test, the interaction of those two variables, a first-degree polynomial of the running variable with different slopes above and below the proficiency threshold, high school and year FE, indicators for EDS, race, and the difference between the standardized math and reading 8th grade test scores. Standard errors are clustered at test score level. Optimal bandwidth is between 4 and 5.

Table A.1: Comparison Between Fall, After Fall and Middle School Samples

	Fall	After Fall	Difference	Middle School	Difference
		(AF)	Fall - AF	(MS)	Fall - MS
	(1)	(2)	(3)	(4)	(5)
Female	0.51	0.47	0.03***	0.51	-0.00
Black	0.26	0.33	-0.07***	0.18	0.08***
Hispanic	0.13	0.15	-0.03***	0.10	0.03***
White	0.54	0.45	0.10***	0.64	-0.10***
EDS	0.51	0.59	-0.08***	0.31	0.20***
Non-Prof. Math 1 test	0.49	0.62	-0.13***	0.24	0.25***
Non-Prof. MS Math test	0.63	0.68	-0.05***	0.23	0.39***
Non-Prof. MS Reading test	0.53	0.57	-0.04***	0.18	0.34***
N	27,997	129,541		76,505	

Note: Table presents sample proportions of variables of interest. Fall sample refers to the students that took the Math 1 class during the fall semester of their freshman year. After Fall sample refers to the students that took Math 1 at any other time after the fall semester of their freshman year. Middle School sample refers to the students that took Math 1 during middle school. The difference columns show the mean difference between the Fall sample and the After Fall sample or Middle School Sample, respectively, for each variable and its significance. Non-Prof.: non-proficiency, MS: middle school, EDS: Economically Disadvantaged Student. *Significant at 10%, **5%, ***1%.

Table A.2: Comparison Between Fall and After Fall Samples

	Fall	After Fall	Difference
Female	0.51	0.49	0.02***
Black	0.24	0.27	-0.03***
Hispanic	0.11	0.12	-0.01***
White	0.58	0.52	0.05***
EDS	0.42	0.46	-0.04***
Non-Prof. English 2 test	0.35	0.34	0.01**
Non-Prof. Middle School Math test	0.35	0.32	0.03***
Non-Prof. Middle School Reading test	0.38	0.35	0.03***
N	72,395	143,057	

Note: Table presents sample proportions of variables of interest. Fall sample refers to the students that took the English 2 class during the fall semester of their sophomore year. After Fall sample refers to the students that took English 2 at any other time after the fall semester of their sophomore year. The difference column shows the mean difference between the two samples for each variable and its significance. Non-Prof.: non-proficiency, EDS: Economically Disadvantaged Student. *Significant at 10%, **5%, ***1%.

Table A.3: Probability of Taking Honors Math 2. ±5 bandwidth

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Female (F)	0.107***	0.099***	0.099***	0.099***	0.098***	0.099***	0.098***	0.101***
	(0.009)	(0.005)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
Non-Prof.	-0.009	-0.006	-0.008	-0.008	-0.008	-0.009	-0.009	-0.009
	(0.007)	(0.008)	(0.008)	(0.007)	(0.007)	(0.006)	(0.006)	(0.007)
F*Non-Prof.	-0.051***	-0.041***	-0.040***	-0.039***	-0.039***	-0.038***	-0.038***	-0.038***
	(0.012)	(0.008)	(0.008)	(0.008)	(0.009)	(0.008)	(0.009)	(0.008)
Math Test MS						0.099***	0.097***	
						(0.009)	(0.008)	
Reading Test MS							0.006	
							(0.007)	
(Math-Reading) Test MS								0.026***
								(0.005)
School FE		✓	✓	✓	✓	✓	✓	✓
Year FE			✓	✓	✓	✓	✓	✓
EDS				✓	✓	✓	✓	✓
Race					✓	✓	✓	✓
Mean	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0.22
R^2	0.05	0.22	0.22	0.23	0.23	0.24	0.24	0.23
N	12,967	12,967	12,967	12,967	12,967	12,322	12,287	12,287

Note: The dependent variable is the same across all specifications: one when taking honors Math 2, zero otherwise. Each specification includes indicator variables for being a female (F), for being non-proficient (Non-Prof) on the Math 1 EOC test, and the interaction of those two variables. Additionally all specifications include a first-degree polynomial of the running variable with different slopes above and below the proficiency threshold. Reading Test MS: standardized middle school reading test score. Math test MS: standardized middle school math test score. (Math-Reading) Test MS: difference between the standardized math and reading middle school test scores. School FE: school fixed effects. EDS: Economically Disadvantaged Student. Standard errors are clustered at test score level.*Significant at 10%, **5%, ***1%.

Table A.4: Probability of Taking Honors Math 2. ±7 bandwidth

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Female (F)	0.125***	0.115***	0.115***	0.115***	0.114***	0.115***	0.114***	0.117***
	(0.013)	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)	(0.012)	(0.011)
Non-Prof.	0.001	-0.002	-0.001	-0.001	-0.002	-0.000	-0.000	-0.002
	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.007)	(0.007)	(0.008)
F*Non-Prof.	-0.075***	-0.064***	-0.063***	-0.063***	-0.062***	-0.061***	-0.061***	-0.062***
	(0.015)	(0.012)	(0.012)	(0.012)	(0.013)	(0.013)	(0.013)	(0.013)
Math Test MS						0.097***	0.095***	
						(0.010)	(0.010)	
Reading Test MS							0.006	
							(0.006)	
(Math-Reading) Test MS								0.025***
								(0.005)
School FE		✓	✓	✓	✓	✓	✓	✓
Year FE			✓	✓	✓	✓	✓	✓
EDS				✓	✓	✓	✓	✓
Race					✓	✓	✓	✓
Mean	0.23	0.23	0.23	0.23	0.23	0.24	0.24	0.24
R^2	0.08	0.23	0.23	0.24	0.24	0.25	0.25	0.24
N	17,339	17,339	17,339	17,339	17,339	16,465	16,417	16,417

Note: The dependent variable is the same across all specifications: one when taking honors Math 2, zero otherwise. Each specification includes indicator variables for being a female (F), for being non-proficient (Non-Prof) on the Math 1 EOC test, and the interaction of those two variables. Additionally all specifications include a first-degree polynomial of the running variable with different slopes above and below the proficiency threshold. Reading Test MS: standardized middle school reading test score. Math test MS: standardized middle school math test score. (Math-Reading) Test MS: difference between the standardized math and reading middle school test scores. School FE: school fixed effects. EDS: Economically Disadvantaged Student. Standard errors are clustered at test score level.*Significant at 10%, **5%, ***1%.

Table A.5: Probability of Taking Honors Math 2.

	(1)	(2)	(3)	(4)	(5)
Female (F)	0.111***	0.113***	0.106***	0.110***	0.105***
	(0.017)	(0.010)	(0.009)	(0.011)	(0.008)
Non-Prof.	-0.014	-0.011	-0.005	-0.011*	-0.008
	(0.008)	(0.006)	(0.006)	(0.006)	(0.005)
Z	-0.064***	0.029	0.040***	0.038	-0.001
	(0.006)	(0.019)	(0.008)	(0.043)	(0.040)
F*Non-Prof.	-0.050**	-0.049***	-0.048***	-0.045***	-0.048***
	(0.021)	(0.013)	(0.010)	(0.013)	(0.009)
F*Z	-0.007	-0.023	0.019	-0.077	0.057
	(0.016)	(0.016)	(0.023)	(0.061)	(0.051)
Non-Prof.*Z	0.013	0.011	-0.016	0.041	-0.074
	(0.009)	(0.016)	(0.014)	(0.051)	(0.095)
F*Non-Prof.*Z	0.002	0.002	-0.010	0.021	0.089
	(0.025)	(0.020)	(0.039)	(0.071)	(0.143)
Z	EDS	Black	Hispanic	Bottom 10% in Middle School Math Test	Top 10% in Middle School Math Test
Mean	0.23	0.23	0.23	0.23	0.23
R^2	0.25	0.25	0.25	0.25	0.25
N	14,184	14,184	14,184	14,184	14,184

Note: The dependent variable is the same across all specifications: one when taking honors math 2 honors, zero otherwise. Each specification includes a dummy variable for being a female (F), for being non-proficient (Non-Prof) in the Math 1 EOC test and a variable Z that changes across columns as indicated by the row Z in the second section of table, all the possible interactions between these three variables and a triple interaction. Additionally, all specifications include a first-degree polynomial of the running variable with different slopes above and below the proficiency threshold. All columns include controls for school and year fixed effects, EDS, race. EDS: Economically Disadvantaged Student. Top (Bottom) 10% refers to a variable equal to one when the student scored in the top (bottom) 10% of the score distribution on their middle school math test. Standard errors clustered at test score level.*Significant at 10%, **5%, ***1%.

Table A.6: Probability of Taking Advanced English 3. ±4 bandwidth

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Female (F)	0.112***	0.100***	0.100***	0.106***	0.107***	0.121***	0.138***	0.115***
	(0.007)	(0.007)	(0.007)	(0.006)	(0.006)	(0.006)	(0.005)	(0.005)
Non-Prof.	0.006	0.014	0.014	0.012	0.012	0.011	0.015*	0.013
	(0.010)	(0.010)	(0.010)	(0.010)	(0.009)	(0.011)	(0.008)	(0.010)
F*Non-Prof.	-0.015	-0.020	-0.020	-0.018	-0.019	-0.023	-0.021	-0.022
	(0.012)	(0.012)	(0.012)	(0.011)	(0.011)	(0.013)	(0.013)	(0.014)
Reading Test MS						0.118***	0.070***	
						(0.006)	(0.005)	
Math Test MS							0.125***	
							(0.005)	
(Math-Reading) Test MS								0.052***
								(0.004)
School FE		✓	✓	✓	✓	✓	✓	✓
Year FE			✓	✓	✓	✓	✓	✓
EDS				✓	✓	✓	✓	✓
Race					✓	✓	✓	✓
Mean	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.40
R^2	0.04	0.20	0.20	0.22	0.22	0.24	0.26	0.23
N	21,472	21,472	21,472	21,472	21,472	19,966	19,935	19,935

Note: The dependent variable is the same across all specifications: one when taking an advanced English 3 class, zero otherwise. Each specification includes indicator variables for being a female (F), for being non-proficient (Non-Prof) on the English 2 EOC test; and the interaction of those two variables. Additionally all specifications include a first-degree polynomial of the running variable with different slopes above and below the proficiency threshold. Reading Test MS: standardized middle school reading test score. Math test MS: standardized middle school math test score. (Math-Reading) Test MS: difference between the standardized math and English middle school test scores. School FE: school fixed effects. EDS: Economically Disadvantaged Student. Standard errors are clustered at test score level.*Significant at 10%, **5%, ***1%.

Table A.7: Probability of Taking Advanced English 3. ±6 bandwidth

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Female (F)	0.111***	0.099***	0.099***	0.105***	0.105***	0.118***	0.135***	0.113***
	(0.006)	(0.006)	(0.006)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
Non-Prof.	0.011	0.012	0.012	0.010	0.011	0.011	0.013**	0.011
	(0.009)	(0.008)	(0.008)	(0.008)	(0.007)	(0.008)	(0.006)	(0.009)
F*Non-Prof.	-0.026**	-0.028**	-0.028**	-0.025**	-0.026**	-0.029**	-0.027**	-0.029**
	(0.012)	(0.010)	(0.010)	(0.010)	(0.010)	(0.011)	(0.012)	(0.012)
Reading Test MS						0.116***	0.067***	
						(0.005)	(0.004)	
Math Test MS							0.129***	
							(0.007)	
(Math-Reading) Test MS								0.055***
								(0.006)
School FE		✓	✓	✓	✓	✓	✓	✓
Year FE			✓	✓	✓	✓	✓	✓
EDS				✓	✓	✓	✓	✓
Race					✓	✓	✓	✓
Mean	0.42	0.42	0.42	0.42	0.42	0.42	0.42	0.42
R^2	0.08	0.22	0.22	0.24	0.24	0.26	0.29	0.25
N	31,495	31,495	31,495	31,495	31,495	29,243	29,199	29,199

Note: The dependent variable is the same across all specifications: one when taking an advanced English 3 class, zero otherwise. Each specification includes indicator variables for being a female (F), for being non-proficient (Non-Prof) on the English 2 EOC test; and the interaction of those two variables. Additionally all specifications include a first-degree polynomial of the running variable with different slopes above and below the proficiency threshold. Reading Test MS: standardized middle school reading test score. Math test MS: standardized middle school math test score. (Math-Reading) Test MS: difference between the standardized math and English middle school test scores. School FE: school fixed effects. EDS: Economically Disadvantaged Student. Standard errors are clustered at test score level.*Significant at 10%, **5%, ***1%.

Table A.8: Probability of Taking Advanced English 3

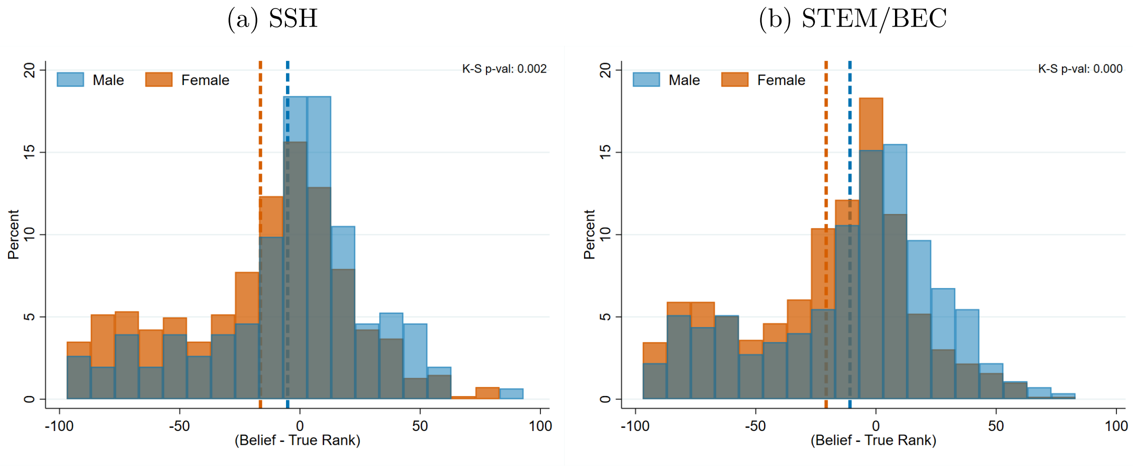
	(1)	(2)	(3)	(4)	(5)
Female (F)	0.143***	0.125***	0.120***	0.120***	0.118***
	(0.006)	(0.007)	(0.005)	(0.006)	(0.005)
Non-Prof.	-0.015	0.012	0.007	0.007	0.010
	(0.014)	(0.010)	(0.011)	(0.009)	(0.009)
Z	-0.135***	0.017	0.015	0.042	-0.045**
	(0.014)	(0.011)	(0.009)	(0.027)	(0.018)
F*Non-Prof.	-0.026	-0.032**	-0.024*	-0.023	-0.024*
	(0.017)	(0.012)	(0.012)	(0.015)	(0.012)
F*Z	-0.052***	-0.031**	-0.018	-0.060***	0.016
	(0.010)	(0.010)	(0.018)	(0.011)	(0.016)
Non-Prof.*Z	0.048**	-0.015	0.012	0.010	-0.112*
	(0.017)	(0.011)	(0.023)	(0.037)	(0.056)
F*Non-Prof.*Z	0.006	0.032*	-0.003	0.028	0.124
	(0.015)	(0.015)	(0.033)	(0.035)	(0.160)
Z	EDS	Black	Hispanic	Bottom 10% in Middle School Reading Test	Top 10% in Middle School Reading Test
Mean	0.42	0.42	0.42	0.42	0.42
R^2	0.25	0.25	0.25	0.25	0.25
N	25,175	25,175	25,175	25,175	25,175

Note: The dependent variable is the same across all specifications: one when taking an advanced English 3 class, zero otherwise. Each specification includes an indicator variable for being a female (F), for being non-proficient (Non-Prof) on the English 2 EOC test and a variable Z that changes across columns as indicated by the row Z in the second section of table, as well as all the possible interactions between these three variables and a triple interaction. Additionally all specifications include a first-degree polynomial of the running variable with different slopes above and below the proficiency threshold. All columns include controls for school and year fixed effects, EDS, race. EDS: Economically Disadvantaged Student. Top (Bottom) 10% refers to a variable equal to one when the student scored in the top (bottom) 10% of the score distribution in their middle school reading test. Standard errors are clustered at test score level.*Significant at 10%, **5%, ***1%.

APPENDIX B

ADDITIONAL FIGURES AND TABLES, CHAPTER 2

Figure B.1: Ability Over/Under Confidence, by Majors


(a) SSH

(b) STEM/BEC

Note: Histogram, by gender and major, of the difference between participants' beliefs about their rank in their reported major and their "true" rank in that major based on reported cumulative GPA. Dashed lines represent the mean of each respective distribution. K-S p-val: p-value from a Kolmogorov-Smirnov test for the equality of the distributions.

Table B.1: Relationship between Gender Gaps in WTP for GPA and Pontential Mechanisms

	(1)	(2)	(3)	(4)
Female	3,057**	2,905**	2,664*	1,437
	(1,440)	(1,439)	(1,528)	(2,029)
Error in Beliefs about Own Ability		-18	-17	-15
		(20)	(20)	(20)
Error in Beliefs about GPA at Graduation			1,770	174
			(2,456)	(2,507)
Anticipated Discrimination SSH				-300
				(545)
Anticipated Discrimination STEM/BEC				636
				(742)
Belief GPA Threshold SSH				-464
				(2,062)
Belief GPA Threshold STEM/BEC				5,363**
				(2,457)
Mean	8,309	8,309	8,309	8,309
N	1,192	1,192	1,192	1,192

Note: Outcome variable is WTP for an extra point in av. GPA at graduation. All columns control for household income, parents education, SAT/ACT, school year, honors, minority, and major. Bootstrapped standard errors reported in parentheses. *Significant at 10%, **5%, ***1%.

APPENDIX C

EXISTENCE OF θ_I^*, CHAPTER 2

Student i stays in major S iff

$$\begin{aligned}
&U_g^S(\theta_i) \geq U_g^N(\theta_i) \\
&\iff P'(\theta_i)[1 - F_h(\tilde{\theta}_g^S)] + (1 - P'(\theta_i))[1 - F_l(\tilde{\theta}_g^S)] - \delta^S c_i \\
&\qquad \geq vP'(\theta_i)[1 - F_h(\tilde{\theta}_g^N)] + v(1 - P'(\theta_i))[1 - F_l(\tilde{\theta}_g^N)] - \delta^N c_i \\
&\iff P'(\theta_i) \geq \frac{c_i(\delta^N - \delta^S) + [F_l(\tilde{\theta}_g^N) - F_l(\widetilde{\theta_g^S})]}{v[F_l(\tilde{\theta}_g^N) - F_h(\tilde{\theta}_g^N)] - [F_l(\tilde{\theta}_g^S) - F_h(\tilde{\theta}_g^S)]}
\end{aligned}$$

Let $\Xi_i = \frac{c_i(\delta^N - \delta^S) + [F_l(\tilde{\theta}^N) - F_l(\tilde{\theta}^S)]}{v[F_l(\tilde{\theta}^N) - F_h(\tilde{\theta}^N)] - [F_l(\tilde{\theta}^S) - F_h(\tilde{\theta}^S)]}$

By MLRP $P'(\theta_i)$ is continuous and increasing in $[0, 1]$, then $P'(\theta_i) \geq \Xi_i$ holds if and only if $\theta_i \geq \theta_i^*$, where the threshold θ_i^* is determined as follows.

- If $P'(0) \geq \Xi_i$ then $\theta_i^* = 0$
- If $P'(1) \leq \Xi_i$ then $\theta_i^* = 1$
- If $P'(0) < \Xi_i$ and $P'(1) > \Xi_i$, by the Intermediate Value Theorem $\exists \theta_i^* \in (0, 1)$ s.t.
$$P'(\theta_i^*) = \Xi_i \tag{C.1}$$

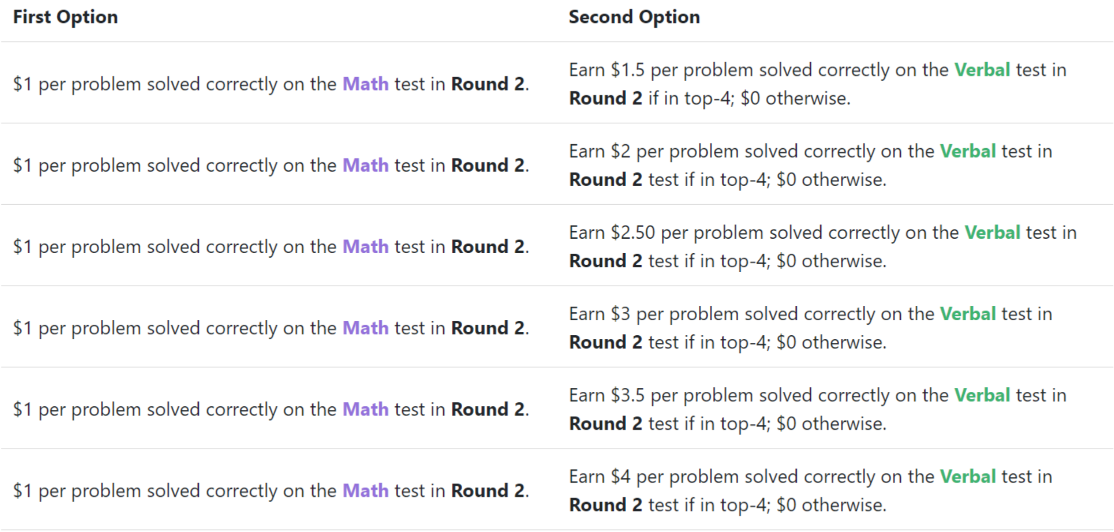
The first two cases imply the everyone stays in S, or switches to N, respectively. The third case is more intuitive,

- if $\theta_i \geq \theta_i^* \Rightarrow P'(\theta_i) \geq \Xi_i \Rightarrow U_g^S(\theta_i) \geq U_g^N(\theta_i)$, individual i stays in S
- if $\theta_i < \theta_i^* \Rightarrow P'(\theta_i) < \Xi_i \Rightarrow U_g^S(\theta_i) < U_g^N(\theta_i)$, individual i changes to N

APPENDIX D

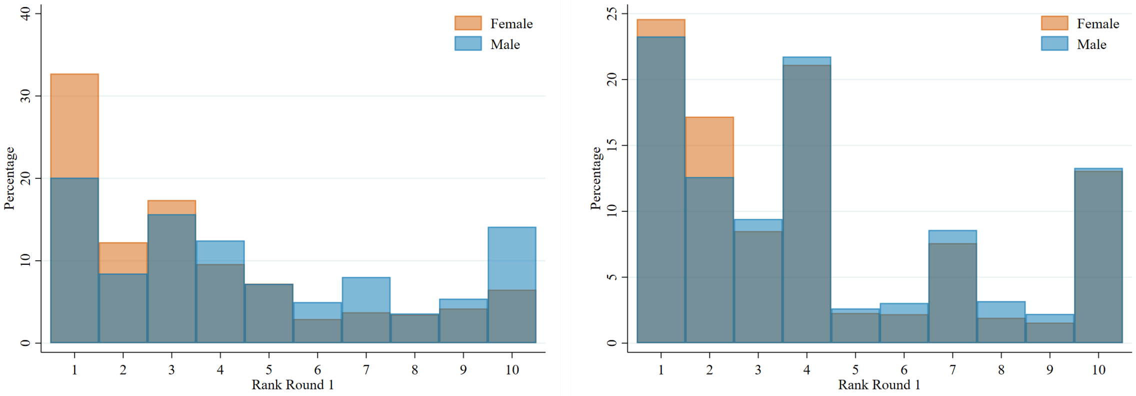
ADDITIONAL FIGURES AND TABLES, CHAPTER 3

Figure D.1: Price Lists for Round 2 Payments, Verbal



First Option	**Second Option**
$1 per problem solved correctly on the **Math** test in **Round 2**.	Earn $1.5 per problem solved correctly on the **Verbal** test in **Round 2** if in top-4; $0 otherwise.
$1 per problem solved correctly on the **Math** test in **Round 2**.	Earn $2 per problem solved correctly on the **Verbal** test in **Round 2** test if in top-4; $0 otherwise.
$1 per problem solved correctly on the **Math** test in **Round 2**.	Earn $2.50 per problem solved correctly on the **Verbal** test in **Round 2** test if in top-4; $0 otherwise.
$1 per problem solved correctly on the **Math** test in **Round 2**.	Earn $3 per problem solved correctly on the **Verbal** test in **Round 2** test if in top-4; $0 otherwise.
$1 per problem solved correctly on the **Math** test in **Round 2**.	Earn $3.5 per problem solved correctly on the **Verbal** test in **Round 2** test if in top-4; $0 otherwise.
$1 per problem solved correctly on the **Math** test in **Round 2**.	Earn $4 per problem solved correctly on the **Verbal** test in **Round 2** test if in top-4; $0 otherwise.

Figure D.2: Round 1 Actual Rank Histograms

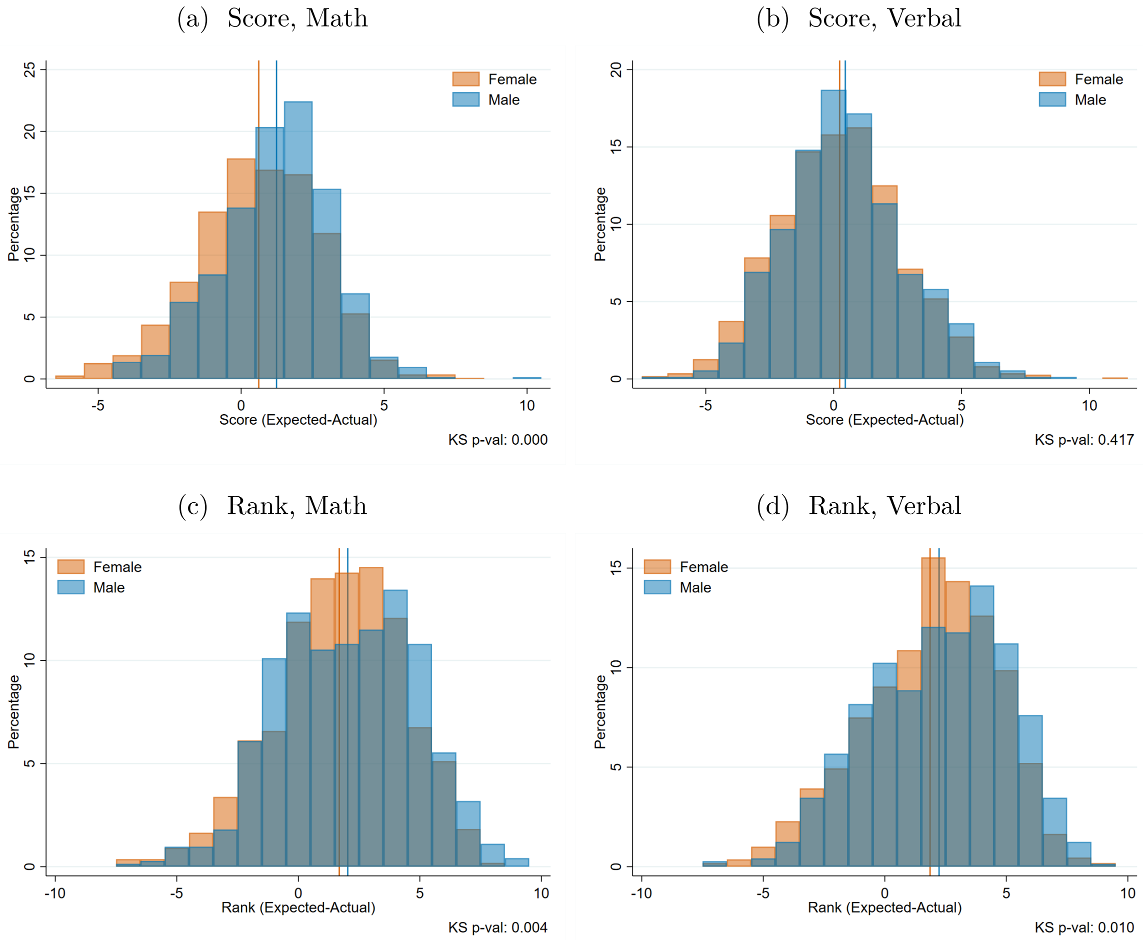


(a) Math

(b) Verbal

Note: Best ranking is ten, worst is one.

Figure D.3: Relative and Absolute Overconfidence by Gender



(a) Score, Math

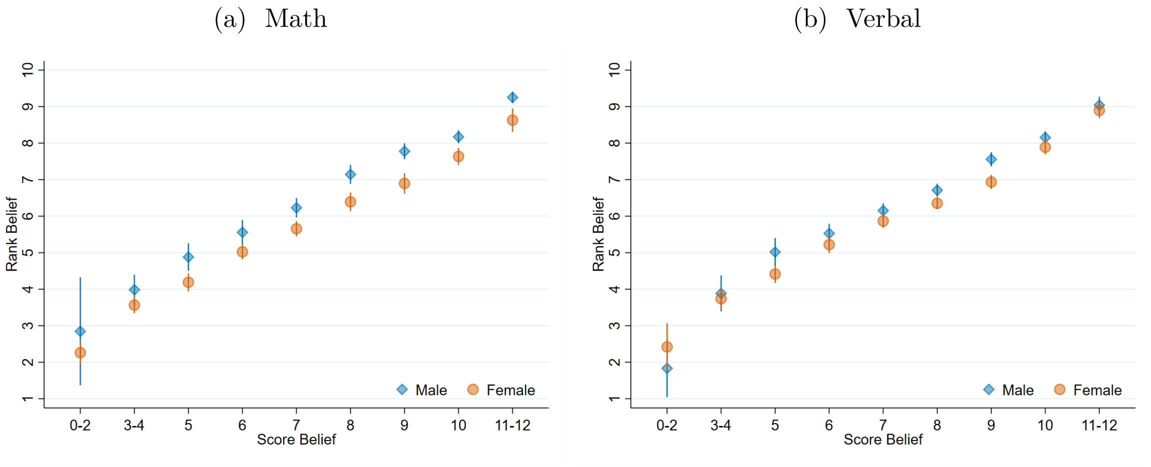
(b) Score, Verbal

(c) Rank, Math

(d) Rank, Verbal

Note: Panel (a) and (b) are histograms of the difference between the initial expected number of correct answers and the actual number of correct answers in each Round 1 quiz, respectively. Panel (c) and (d) are histograms of the difference between the initial expected rank and the actual rank of participants each Round 1 quiz. Vertical lines at the means for each gender. KS p-val is the p-value of a Kolmogorov-Smirnov tests of the equality of distributions.

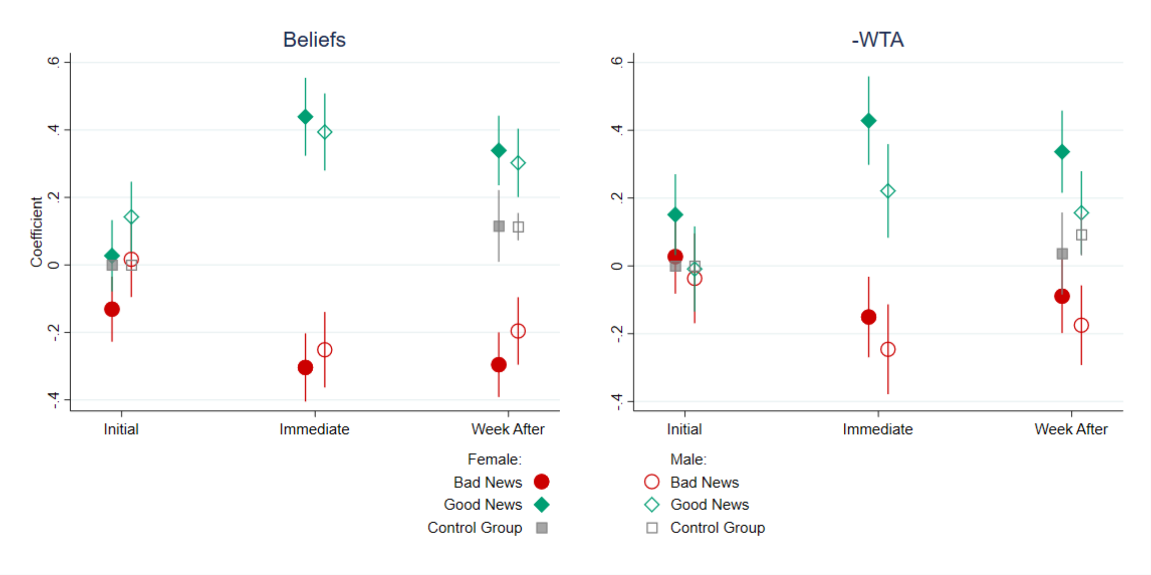
Figure D.4: Rank Belief by Score Belief


(a) Math

(b) Verbal

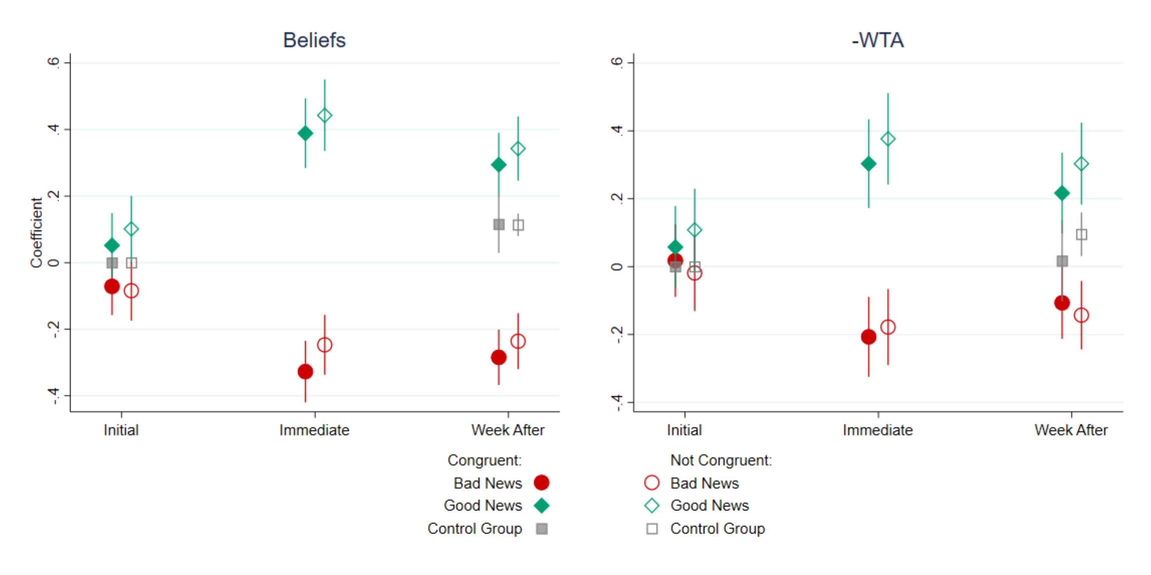
Note: Markers represent the mean of the rank beliefs by each score level on the x-axis. The spikes represent 95% confidence intervals.

Figure D.5: Levels over Time by Gender, Accurate Subsample



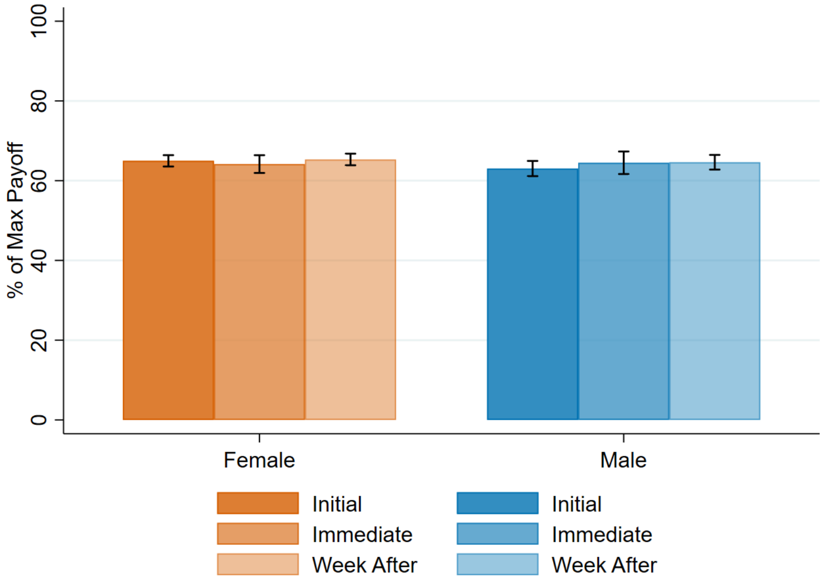
Note: Markers represent the coefficient on good and bad news, and control group at different stages for each gender from a regression that pools all stages for all the participants that accurately remember their feedback for that domain. The outcome, indicated at the top of each panel, is regressed on indicator variables for initial good news male and female, immediate good news male and female, immediate bad news male and female, week after good news male and female, week after bad news male and female, initial control female and week after control group male and female (i.e. the omitted category is initial control male), $\boldsymbol{Y}$: relative and absolute performance controls, an indicator variable for the difficulty level of the first round tests and all the controls in $\boldsymbol{X}$: family income, indicators for each parent attending college, a nonwhite indicator, ACT scores, high school rank, indicator for attending high school in the U.S., honors student indicator, school year (freshman, sophomore, junior or senior), a measure of risk aversion, an indicator for taking the math quiz first, domain indicator, and an immediate group indicator. Errors clustered at individual level. Beliefs is a composite variable that aggregates the three different measures of beliefs elicited, it is standardized with mean zero and standard deviation one at every stage. -WTA is the negative of the WTA standardized per stage. The spikes represent 90% confidence intervals.

Figure D.6: Levels over Time by Domain Congruence, Accurate Subsample



Note: Markers represent the coefficient on good and bad news, and control group at different stages for congruent and incongruent domains from a regression that pools all stages for all the participants that accurately remember their feedback for that domain. The outcome, indicated at the top of each panel, is regressed on indicator variables for initial good news congruent and incongruent, immediate good news congruent and incongruent, immediate bad news congruent and incongruent, week after good news congruent and incongruent, week after bad news congruent and incongruent, initial control not congruent and week after control group congruent and incongruent (i.e. the omitted category is initial control congruent), $\boldsymbol{Y}$: relative and absolute performance controls, an indicator variable for the difficulty level of the first round tests and all the controls in $\boldsymbol{X}$: gender, family income, indicators for each parent attending college, a nonwhite indicator, ACT scores, high school rank, indicator for attending high school in the U.S., honors student indicator, school year (freshman, sophomore, junior or senior), a measure of risk aversion, an indicator for taking the math quiz first, and an immediate group indicator. Errors clustered at individual level. Beliefs is a composite variable that aggregates the three different measures of beliefs elicited, it is standardized with mean zero and standard deviation one at every stage. -WTA is the negative of the WTA standardized per stage. The spikes represent 90% confidence intervals.

Figure D.7: Expected Payoff as Percent of Maximum Achievable Payoff by Gender



Note: Expected payoff and maximum achievable payoff are calculated as follows: we randomly select a row from the price lists from each of the three stages of the experiment (1,000 times for each stage), and calculate earnings based on the payoff-maximizing choice (i.e., maximum achievable payoffs) and the observed choice (i.e., expected payoffs). Then, we average over the realized earnings.

Table D.1: Attrition Rate

	All	Treated
	(1)	(2)
Female (F)	-0.003	-0.005
	(0.030)	(0.028)
Non-Immediate	-0.007	
	(0.027)	
Non-Immediate*F	0.033	
	(0.036)	
Immediate	0.003	
	(0.028)	
Immediate*F	0.013	
	(0.037)	
Bad News$_{Verbal}$		0.016
		(0.027)
Bad News$_{Verbal}$*F		0.020
		(0.033)
Bad News$_{Math}$		0.033
		(0.026)
Bad News$_{Math}$*F		0.017
		(0.033)
Score$_{Verbal}$	-0.001	0.000
	(0.003)	(0.004)
Score$_{Math}$	-0.003	0.003
	(0.005)	(0.006)
Initial Belief$_{Verbal}$	0.022*	0.021
	(0.012)	(0.015)
Initial Belief$_{Verbal}$*F	-0.016	-0.020
	(0.015)	(0.019)
Initial Belief$_{Math}$	-0.007	-0.014
	(0.012)	(0.014)
Initial Belief$_{Math}$*F	0.016	0.027
	(0.015)	(0.018)
Income	0.000	0.000
	(0.000)	(0.000)
Minority	-0.008	-0.014
	(0.014)	(0.016)
Math First	-0.019	-0.013
	(0.014)	(0.015)
Honors	-0.083***	-0.087***
	(0.017)	(0.019)
US HS	0.027	0.017
	(0.043)	(0.051)
HS Rank	0.001*	0.001
	(0.000)	(0.000)
ACT	-0.005**	-0.002
	(0.002)	(0.003)
Father College	-0.018	-0.020
	(0.018)	(0.021)
Mother College	-0.006	-0.001
	(0.017)	(0.020)
Mean	0.107	0.109
R2	0.058	0.053
N	2,008	1,612

Note: Outcome variable is an indicator variable equal to one if individual did not participante in Session 2. Standard errors reported in parentheses. *Significant at 10%, **5%, ***1%.

Table D.2: Sample Compared to ASU Population

	Experiment			ASU			P-value[c]
	Female	Male	Gender Diff.	Female	Male	Gender Diff.	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Black	0.04	0.03	0.01	0.03	0.02	0.01	0.717
White	0.67	0.68	-0.02	0.54	0.55	-0.01	0.683
Hispanic	0.17	0.13	0.04	0.22	0.19	0.03	0.666
First Generation[a]	0.27	0.20	0.07	0.24	0.19	0.05	0.300
Family Income[b]	104	118	-14	121	134	-13	0.789
Freshman	0.28	0.23	0.05	0.27	0.26	0.01	0.093
Sophomore	0.23	0.30	-0.06	0.25	0.24	0.01	0.001
Junior	0.25	0.25	-0.01	0.22	0.23	-0.01	0.919
Senior	0.23	0.22	0.02	0.26	0.28	-0.02	0.121
ACT	28.74	29.81	-1.06	26.31	27.63	-1.31	0.315
Sample Size	1,094	722		19,199	20,036		0.000[d]

Note: ASU data includes everyone taking at least one class for credit during the Spring semester of 2018 and attended ASU as their first full-time university. ASU data is weighted such that the proportion of honors students is the same as in our experimental sample (59%). Income and first generation variables for the ASU data are constructed with the data of the first available year, which it is not the first year of college for most of the sample.

[a] Students with no parent with a college degree.

[b] Family income in thousands of dollars.

[c] P-value for whether the gender differences in the experiment sample and the ASU population are different.

[d] P-value for the difference in females proportion between the experiment sample and ASU population.

Table D.3: Effect of Priors Restricted to Non-Extreme Ranks

	Initial	Immediately After		Week After	
	(1)	(2)	(3)	(4)	(5)
Panel A: Beliefs					
Bad News$_D$	0.003	-0.659***	-0.656***	-0.460***	-0.462***
	(0.072)	(0.094)	(0.052)	(0.066)	(0.040)
Bad News$_D$*Female	-0.365***	-0.278***	-0.049	-0.328***	-0.062**
	(0.052)	(0.070)	(0.039)	(0.048)	(0.030)
Good News$_D$*Female	-0.334***	-0.182**	0.040	-0.166***	0.078**
	(0.060)	(0.080)	(0.039)	(0.057)	(0.033)
Bad News$_D$*Congruent	0.204***	0.160***	0.005	0.154***	0.006
	(0.046)	(0.056)	(0.030)	(0.040)	(0.025)
Good News$_D$*Congruent	0.180***	0.192***	0.004	0.141***	0.010
	(0.056)	(0.065)	(0.035)	(0.051)	(0.030)
Prior Beliefs			✓		✓
Mean	0.06	0.06	0.06	0.07	0.07
R2	0.27	0.42	0.83	0.36	0.76
Clusters	1,242	589	589	1,242	1,242
Obs.	1,820	864	864	1,820	1,820
Panel B: - WTA					
Bad News$_D$	0.082	-0.472***	-0.473***	-0.327***	-0.369***
	(0.093)	(0.129)	(0.082)	(0.091)	(0.072)
Bad News$_D$*Female	-0.200***	-0.195**	0.004	-0.235***	-0.110**
	(0.061)	(0.089)	(0.061)	(0.064)	(0.052)
Good News$_D$*Female	-0.081	0.064	0.079	-0.027	0.010
	(0.077)	(0.101)	(0.061)	(0.075)	(0.056)
Bad News$_D$*Congruent	0.170***	0.140*	0.060	0.214***	0.118***
	(0.063)	(0.081)	(0.051)	(0.057)	(0.045)
Good News$_D$*Congruent	0.198***	0.136	0.013	0.098	-0.017
	(0.073)	(0.096)	(0.057)	(0.070)	(0.048)
Prior WTA			✓		✓
Mean	0.06	0.05	0.05	0.04	0.04
R2	0.06	0.17	0.66	0.11	0.47
Clusters	1,132	544	544	1,142	1,142
Obs.	1,647	795	795	1,659	1,659

Note: Sample restricted to participants with true ranks between 2 and 9 in the Round 1 quizzes. Outcome variable in Panel A is a composite variable that aggregates the three different measures of beliefs elicited, it is standardized with mean zero and standard deviation one at every stage. Outcome variable in Panel B is the negative of the WTA standarized per stage. Outcomes are regressed on an indicator for bad news, and interactions of good and bad news with female and congruent (i.e. omitted category is males who receive good news in gender-incongruent domains). All specifications control for $\boldsymbol{Y}$: relative and absolute performance controls, and an indicator variable for the difficulty level of the first round tests; and $\boldsymbol{X}$: family income, indicators for each parent attending college, a nonwhite indicator, ACT scores, high school rank, indicator for attending high school in the U.S., honors student indicator, school year (freshman, sophomore, junior or senior), a measure of risk aversion, an indicator for taking the math quiz first and an indicator for being in Immediate group. D denotes an observation associated with a given domain, and -D denotes the other domain. Errors clustered at individual level. Standard errors reported in parentheses. *Significant at 10%, **5%, ***1%.

Table D.4: Proportion of Participants that Made the Right Choice, Initial

	Chose Math		Chose Verbal	
	Female	Male	Female	Male
	(1)	(2)	(3)	(4)
$4.0 per-math question	0.20	0.25	0.91	0.84
$3.5 per-math question	0.22	0.26	0.90	0.84
$3.0 per-math question	0.22	0.26	0.89	0.82
$2.5 per-math question	0.24	0.30	0.87	0.83
$2.0 per-math question	0.24	0.31	0.86	0.81
$1.5 per-math question	0.22	0.30	0.86	0.79
$1.0 math vs $1.0 verbal	0.52	0.60	0.73	0.72
$1.5 per-verbal question	0.64	0.63	0.35	0.41
$2.0 per-verbal question	0.65	0.65	0.38	0.45
$2.5 per-verbal question	0.66	0.68	0.38	0.45
$3.0 per-verbal question	0.69	0.71	0.40	0.43
$3.5 per-verbal question	0.72	0.74	0.40	0.43
$4.0 per-verbal question	0.74	0.70	0.39	0.40

Note: Correct choice means that participants chose the option that gives a higher payoff given their performance in Round 2. Rows 1-6 report the proportions for the decisions from Figure 3.2, the price list for the competitive payment scheme in math vs $1 verbal. Row 7 report the propotiorn for the piece-rate payment scheme (math vs verbal). Rows 8-13 report the proportions for the decisions from Figure D.1, the price list for the competitive payment scheme in verbal vs $1 math.

Table D.5: Relationship Between Initial Beliefs and Choices

	Choosing Math		WTA Math		WTA Verbal	
	Female	Male	Female	Male	Female	Male
	(1)	(2)	(3)	(4)	(5)	(6)
$\text{Beliefs}_{\text{Verbal}}$	-25.82***	-27.88***	24.39***	20.18***	-66.16***	-67.42***
	(1.47)	(2.12)	(4.08)	(4.69)	(3.87)	(5.08)
$\text{Beliefs}_{\text{Math}}$	35.48***	38.32***	-73.09***	-88.43***	38.38***	28.95***
	(1.63)	(2.05)	(4.26)	(4.74)	(4.56)	(5.78)
Performance Controls	✓	✓	✓	✓	✓	✓
F-test[a]	0.058	0.361	0.082	0.021	0.478	0.121
Mean	39.79	55.08	336.51	285.69	295.28	295.11
R2	0.48	0.49	0.39	0.47	0.33	0.34
N	872	581	767	531	763	542

Note: Outcome variable in column (1) and (2) is an indicator variable equal to one when participant chooses to be compensated for math in Round 2, prior to receiving feedback. Outcome variable in columns (3)-(6) is the initial WTA in cents. Beliefs for each subject is a composite variable that aggregates the three different measures of beliefs elicited before feedback, it is standardized with mean zero and standard deviation one at every stage. The higher the measure the more optimistic the beliefs. All specifications control for performance: score and rank in both domains. Standard errors reported in parentheses. *Significant at 10%, **5%, ***1%.

[a] P-value from F-test for joint significance of the performance controls.

Table D.6: Distribution of Feedback Combinations by Gender

		Math			
		Female		Male	
		Good	Bad	Good	Bad
Verbal	Good	12.50	21.56	20.65	17.90
	Bad	16.17	49.77	23.06	38.38

Table D.7: Bayesian Posterior for Good and Bad News for Different Priors

	Prior	Posterior Good News	Posterior Bad News
	(1)	(2)	(3)
High tightness rank 3	3.00	3.20	2.94
Low tightness rank 3	3.00	4.00	2.71
High tightness rank 6	6.00	6.08	5.90
Low tightness rank 6	6.00	6.40	5.50
High tightness rank 8	8.00	8.06	7.80
Low tightness rank 8	8.00	8.29	7.00

Note: High tightness means that 60% of the mass is in the rank indicated in the row and the other 40% is uniformly distributed between one rank above and one below. Low tightness means that 20% of the mass is in the rank indicated in the row and the other 80% is uniformly distributed between two ranks above and below.

Table D.8: Direction of Change in Rank Beliefs by Feedback Type and Gender

	Female				Male			
	Immediate		Week After		Immediate		Week After	
	Good News	Bad News	Good News	Bad News	Good News	Bad News	Good News	Bad News
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Math								
Adjusted Up	0.52	0.08	0.45	0.20	0.31	0.14	0.30	0.19
No Change	0.41	0.37	0.37	0.37	0.59	0.26	0.51	0.39
Adjusted Down	0.08	0.55	0.18	0.43	0.10	0.60	0.20	0.42
Panel B: Verbal								
Adjusted Up	0.46	0.07	0.43	0.19	0.30	0.11	0.32	0.19
No Change	0.43	0.33	0.41	0.38	0.61	0.32	0.50	0.39
Adjusted Down	0.11	0.60	0.16	0.43	0.09	0.57	0.18	0.42

Note: The table reports the proportion of participants that updated up, down or did not change their rank belief guess after receiving feedback. The immediate change (columns 1,2,5,6) is the calculated only for participants in the immediate group as the immediate measures minus the initial. The week after change (columns 3,4,7,8) is calculated as week after measure minus initial measure. The shaded cells represent proportion of participants for which the direction of the change in beliefs is what we would expect given the type of feedback.

Table D.9: Effect of Priors, Monotonic Decision Makers Subsample

	Initial	Immediately After		Week After	
	(1)	(2)	(3)	(4)	(5)
Panel A: Beliefs					
Bad News$_D$	0.003	-0.666***	-0.608***	-0.442***	-0.444***
	(0.068)	(0.090)	(0.048)	(0.063)	(0.036)
Bad News$_D$*Female	-0.372***	-0.282***	-0.072**	-0.345***	-0.068***
	(0.048)	(0.062)	(0.035)	(0.043)	(0.026)
Good News$_D$*Female	-0.312***	-0.165**	0.053	-0.148***	0.084***
	(0.053)	(0.071)	(0.032)	(0.049)	(0.026)
Bad News$_D$*Congruent	0.177***	0.090*	-0.022	0.113***	-0.018
	(0.041)	(0.049)	(0.024)	(0.034)	(0.019)
Good News$_D$*Congruent	0.147***	0.115**	-0.011	0.119***	0.009
	(0.048)	(0.057)	(0.028)	(0.044)	(0.023)
Prior Beliefs			✓		✓
Mean	0.00	-0.00	-0.00	0.00	0.00
R2	0.41	0.55	0.88	0.52	0.84
Clusters	1,145	551	551	1,145	1,145
Obs.	2,290	1,102	1,102	2,290	2,290
Panel B: - WTA					
Bad News$_D$	-0.031	-0.562***	-0.395***	-0.342***	-0.322***
	(0.084)	(0.116)	(0.072)	(0.081)	(0.063)
Bad News$_D$*Female	-0.191***	-0.192***	-0.046	-0.241***	-0.120***
	(0.051)	(0.073)	(0.050)	(0.051)	(0.043)
Good News$_D$*Female	-0.075	-0.057	0.023	-0.069	-0.021
	(0.062)	(0.083)	(0.047)	(0.062)	(0.044)
Bad News$_D$*Congruent	0.246***	0.161**	0.016	0.204***	0.048
	(0.052)	(0.067)	(0.041)	(0.047)	(0.035)
Good News$_D$*Congruent	0.166***	0.138*	0.049	0.109**	0.004
	(0.058)	(0.078)	(0.043)	(0.055)	(0.037)
Prior WTA			✓		✓
Mean	0.00	-0.00	-0.00	0.00	0.00
R2	0.17	0.27	0.72	0.23	0.57
Clusters	1,145	551	551	1,145	1,145
Obs.	2,290	1,102	1,102	2,290	2,290

Note: The sample excludes participants for which at some point (initial, immediate or week later) for at least one of the domains, is not possible to calculate the WTA because their price list decisions were not monotonic. Outcome variable in Panel A is a composite variable that aggregates the three different measures of beliefs elicited, it is standardized with mean zero and standard deviation one at every stage. Outcome variable in Panel B is the negative of the WTA standarized per stage. Outcomes are regressed on an indicator for bad news, and interactions of good and bad news with female and congruent (i.e. omitted category is males who receive good news in gender-incongruent domains). All specifications control for $\boldsymbol{Y}$: relative and absolute performance controls, and an indicator variable for the difficulty level of the first round tests; and $\boldsymbol{X}$: family income, indicators for each parent attending college, a nonwhite indicator, ACT scores, high school rank, indicator for attending high school in the U.S., honors student indicator, school year (freshman, sophomore, junior or senior), a measure of risk aversion, an indicator for taking the math quiz first and an indicator for being in Immediate group. D denotes an observation associated with a given domain, and -D denotes the other domain. Errors clustered at individual level. Standard errors reported in parentheses. *Significant at 10%, **5%, ***1%.

Table D.10: Effect of Type of News on Beliefs and WTA over Time

	Beliefs		-WTA	
	(1)	(2)	(3)	(4)
Initial, Good News$_D$	0.038	0.036	0.061	-0.007
	(0.044)	(0.049)	(0.053)	(0.059)
Initial, Bad News$_D$	-0.048	-0.050	0.018	-0.061
	(0.041)	(0.046)	(0.047)	(0.055)
Immediate, Good News$_D$	0.374***	0.372***	0.318***	0.248***
	(0.047)	(0.053)	(0.058)	(0.064)
Immediate, Bad News$_D$	-0.255***	-0.257***	-0.172***	-0.250***
	(0.043)	(0.047)	(0.051)	(0.057)
Week After, Good News$_D$	0.267***	0.265***	0.226***	0.157***
	(0.043)	(0.048)	(0.053)	(0.058)
Week After, Bad News$_D$	-0.216***	-0.218***	-0.096**	-0.174***
	(0.040)	(0.045)	(0.047)	(0.055)
Week After, Control Group$_D$	0.115***	0.115***	0.057*	0.057*
	(0.016)	(0.016)	(0.032)	(0.032)
Bad News$_{\text{-}D}$		0.003		0.120***
		(0.037)		(0.043)
Mean	-0.00	-0.00	0.00	0.00
R2	0.46	0.46	0.19	0.19
Clusters	1,816	1,816	1,757	1,757
Obs.	8,642	8,642	7,806	7,806

Note: Outcome variable in columns 1 and 2 is a composite variable that aggregates the three different measures of beliefs elicited, it is standardized with mean zero and standard deviation one at every stage. Outcome variable in column 3 and 4 is the negative of the standardized WTA at every stage. The omitted category is initial control group. All specifications control for $\boldsymbol{Y}$: relative and absolute performance controls, and an indicator variable for the difficulty level of the first round tests; and $\boldsymbol{X}$: family income, indicators for each parent attending college, a nonwhite indicator, ACT scores, high school rank, indicator for attending high school in the U.S., honors student indicator, school year (freshman, sophomore, junior or senior), a measure of risk aversion, an indicator for taking the math quiz first, a domain indicator and immediate group indicator. D denotes an observation associated with a given domain, and -D denotes the other domain. Errors clustered at individual level. Standard errors reported in parentheses. *Significant at 10%, **5%, ***1%.

Table D.11: Effect of Type of News and Congruency of the Domain on Beliefs and WTA over Time

	Beliefs	-WTA	Beliefs	-WTA
	(1)	(2)	(3)	(4)
Initial, Good News$_D$, Non-Congruent	0.059	0.076	0.033	0.046
	(0.058)	(0.070)	(0.056)	(0.070)
Initial, Bad News$_D$, Non-Congruent	-0.051	0.004	-0.035	-0.005
	(0.052)	(0.062)	(0.052)	(0.062)
Initial, Good News$_D$, Congruent	0.192***	0.239***	0.207***	0.227***
	(0.057)	(0.071)	(0.059)	(0.071)
Initial, Bad News$_D$, Congruent	0.131**	0.228***	0.099*	0.194***
	(0.052)	(0.064)	(0.053)	(0.064)
Immediate, Good News$_D$, Non-Congruent	0.394***	0.353***	0.367***	0.327***
	(0.063)	(0.079)	(0.063)	(0.077)
Immediate, Bad News$_D$, Non-Congruent	-0.213***	-0.153**	-0.220***	-0.168**
	(0.054)	(0.067)	(0.054)	(0.066)
Immediate, Good News$_D$, Congruent	0.525***	0.472***	0.541***	0.449***
	(0.062)	(0.077)	(0.064)	(0.079)
Immediate, Bad News$_D$, Congruent	-0.125**	0.001	-0.136**	-0.031
	(0.055)	(0.070)	(0.057)	(0.072)
Week After, Good News$_D$, Non-Congruent	0.295***	0.277***	0.259***	0.242***
	(0.057)	(0.070)	(0.056)	(0.070)
Week After, Bad News$_D$, Non-Congruent	-0.192***	-0.117*	-0.181***	-0.111*
	(0.051)	(0.060)	(0.051)	(0.061)
Week After, Good News$_D$, Congruent	0.414***	0.372***	0.439***	0.353***
	(0.056)	(0.070)	(0.058)	(0.071)
Week After, Bad News$_D$, Congruent	-0.067	0.121*	-0.101**	0.072
	(0.050)	(0.063)	(0.051)	(0.064)
Initial, Control Group$_D$, Congruent	0.174***	0.193***	0.152**	0.140*
	(0.054)	(0.072)	(0.060)	(0.074)
Week After, Control Group$_D$, Non-Congruent	0.114***	0.096**	0.117***	0.055
	(0.021)	(0.039)	(0.022)	(0.045)
Week After, Control Group$_D$, Congruent	0.289***	0.209***	0.263***	0.197***
	(0.052)	(0.072)	(0.059)	(0.073)
Mean	-0.00	0.00	-0.00	0.00
R2	0.47	0.20	0.47	0.19
Clusters	1,816	1,757	1,816	1,757
Obs.	8,642	7,806	8,642	7,806

Note: Outcome variable in columns 1 and 3 is a composite variable that aggregates the three different measures of beliefs elicited, it is standardized with mean zero and standard deviation one at every stage. Outcome variable in columns 2 and 4 is the negative of the standardized WTA at every stage. The omitted category is initial control not congruent. All specifications control for $\boldsymbol{Y}$: relative and absolute performance controls, and an indicator variable for the difficulty level of the first round tests; and $\boldsymbol{X}$: gender, family income, indicators for each parent attending college, a nonwhite indicator, ACT scores, high school rank, indicator for attending high school in the U.S., honors student indicator, school year (freshman, sophomore, junior or senior), a measure of risk aversion, an indicator for taking the math quiz first, and immediate group indicator. In columns 1 and 2 congruent is equal to 1 if the domain is congruent with the the individuals gender (i.e. it is one for males when the domain is math, and one for females when the domain is verbal). In columns 3 and 4 congruent is equal to one if the participant believes that their gender has an advantage in that domain, zero otherwise. D denotes an observation associated with a given domain. Errors clustered at individual level. Standard errors reported in parentheses. *Significant at 10%, **5%, ***1%.

Table D.12: Effect of Type of News and Gender on Beliefs and WTA over Time

	Beliefs	-WTA
	(1)	(2)
Initial, Good News$_D$, Male	0.110*	-0.011
	(0.061)	(0.074)
Initial, Bad News$_D$, Male	0.064	-0.014
	(0.061)	(0.069)
Initial, Good News$_D$, Female	-0.222***	-0.115
	(0.062)	(0.070)
Initial, Bad News$_D$, Female	-0.321***	-0.187***
	(0.058)	(0.066)
Immediate, Good News$_D$, Male	0.367***	0.185**
	(0.067)	(0.083)
Immediate, Bad News$_D$, Male	-0.215***	-0.230***
	(0.067)	(0.077)
Immediate, Good News$_D$, Female	0.178***	0.188**
	(0.068)	(0.077)
Immediate, Bad News$_D$, Female	-0.485***	-0.360***
	(0.061)	(0.071)
Week After, Good News$_D$, Male	0.258***	0.134*
	(0.060)	(0.072)
Week After, Bad News$_D$, Male	-0.133**	-0.125*
	(0.060)	(0.070)
Week After, Good News$_D$, Female	0.077	0.068
	(0.061)	(0.071)
Week After, Bad News$_D$, Female	-0.474***	-0.304***
	(0.058)	(0.065)
Initial, Control Group$_D$, Female	-0.208***	-0.230***
	(0.065)	(0.074)
Week After, Control Group$_D$, Male	0.113***	0.092**
	(0.025)	(0.038)
Week After, Control Group$_D$, Female	-0.093	-0.194***
	(0.065)	(0.073)
Mean	-0.00	0.00
R2	0.47	0.20
Clusters	1,816	1,757
Obs.	8,642	7,806

Note: Outcome variable in columns (1) is a composite variable that aggregates the three different measures of beliefs elicited, it is standardized with mean zero and standard deviation one at every stage. Outcome variable in column (2) is the negative of the standardized WTA at every stage. The omitted category is initial control males. All specifications control for $\boldsymbol{Y}$: relative and absolute performance controls, and an indicator variable for the difficulty level of the first round tests; and $\boldsymbol{X}$: gender, family income, indicators for each parent attending college, a nonwhite indicator, ACT scores, high school rank, indicator for attending high school in the U.S., honors student indicator, school year (freshman, sophomore, junior or senior), a measure of risk aversion, an indicator for taking the math quiz first, a domain indicator and immediate group indicator. D denotes an observation associated with a given domain. Errors clustered at individual level. Standard errors reported in parentheses. *Significant at 10%, **5%, ***1%.

Table D.13: Posterior Expected Rank on Bayesian Update

	Immediately After Feedback							
	Math				Verbal			
	Female		Male		Female		Male	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Bayes	0.874***	0.877***	0.882***	0.850***	0.772***	0.700***	0.841***	0.820***
	(0.030)	(0.036)	(0.040)	(0.050)	(0.038)	(0.046)	(0.043)	(0.052)
Good News$_D$	0.145	0.223	-0.203	-0.803	0.247*	-1.025**	-0.077	-0.513
	(0.125)	(0.422)	(0.152)	(0.577)	(0.142)	(0.507)	(0.155)	(0.632)
Good News$_D$*Bayes		-0.013		0.090		0.206***		0.066
		(0.067)		(0.083)		(0.079)		(0.092)
Constant	0.584***	0.568***	0.748***	0.909***	1.050***	1.376***	0.997***	1.102***
	(0.138)	(0.161)	(0.214)	(0.261)	(0.184)	(0.221)	(0.228)	(0.271)
Mean	4.936	4.936	6.062	6.062	5.345	5.345	5.906	5.906
R2	0.774	0.774	0.748	0.749	0.668	0.673	0.698	0.699
N	411	411	278	278	411	411	278	278
Estimated Responsiveness to:								
Good News		0.864		0.940		0.906		0.886
Bad News		0.877		0.850		0.700		0.820
	Week After							
Bayes	0.846***	0.851***	0.840***	0.768***	0.770***	0.713***	0.818***	0.766***
	(0.023)	(0.027)	(0.026)	(0.031)	(0.025)	(0.030)	(0.028)	(0.034)
Good News$_D$	-0.104	0.030	-0.345***	-1.837***	0.067	-1.001***	-0.278***	-1.396***
	(0.097)	(0.338)	(0.101)	(0.383)	(0.095)	(0.340)	(0.100)	(0.416)
Good News$_D$*Bayes		-0.022		0.220***		0.173***		0.166***
		(0.053)		(0.055)		(0.053)		(0.060)
Constant	0.893***	0.870***	1.072***	1.440***	1.245***	1.500***	1.190***	1.459***
	(0.102)	(0.116)	(0.142)	(0.167)	(0.119)	(0.142)	(0.152)	(0.179)
Mean	4.959	4.959	6.118	6.118	5.370	5.370	6.009	6.009
R2	0.720	0.720	0.731	0.738	0.672	0.676	0.692	0.696
N	872	872	581	581	872	872	581	581
Estimated Responsiveness to:								
Good News		0.829		0.988		0.886		0.932
Bad News		0.851		0.768		0.713		0.766

Note: Outcome variable and bayesian update correspond to the expected rank given the probability distributions. D denotes an observation associated with a given domain. Standard errors reported in parentheses. *Significant at 10%, **5%, ***1%.

Table D.14: Male-Female Gap over Time

	Beliefs	-WTA
	(1)	(2)
Initial, Treated Female	-0.347***	-0.135***
	(0.032)	(0.037)
Immediate, Treated Female	-0.262***	-0.094*
	(0.045)	(0.052)
Week After, Treated Female	-0.304***	-0.149***
	(0.031)	(0.037)
Immediate, Treated	-0.068***	-0.052
	(0.026)	(0.032)
Week After, Treated	-0.055***	-0.004
	(0.015)	(0.026)
Initial, Control Female	-0.191***	-0.218***
	(0.065)	(0.073)
Week After, Control Female	-0.188***	-0.273***
	(0.066)	(0.072)
Initial, Control	-0.082	0.013
	(0.057)	(0.063)
Week After, Control	0.031	0.103*
	(0.057)	(0.062)
Mean	-0.00	0.00
R2	0.44	0.18
Clusters	1,816	1,757
Obs.	8,642	7,806

Note: Outcome variable in columns (1) is a composite variable that aggregates the three different measures of beliefs elicited, it is standardized with mean zero and standard deviation one at every stage. Outcome variable in column (2) is the negative of the standardized WTA at every stage. The omitted category is initial treated. All specifications control for $\boldsymbol{Y}$: relative and absolute performance controls, and an indicator variable for the difficulty level of the first round tests; and $\boldsymbol{X}$: family income, indicators for each parent attending college, a non-white indicator, ACT scores, high school rank, indicator for attending high school in the U.S., honors student indicator, school year (freshman, sophomore, junior or senior), a measure of risk aversion, an indicator for taking the math quiz first, and immediate group indicator. Errors clustered at individual level. Standard errors reported in parentheses. *Significant at 10%, **5%, ***1%.

Table D.15: Stereotype Differences over Time

	Beliefs	-WTA	Beliefs	-WTA
	(1)	(2)	(3)	(4)
Initial, Treated Congruent	0.155***	0.194***	0.145***	0.190***
	(0.027)	(0.037)	(0.030)	(0.038)
Immediate, Treated Congruent	0.131***	0.166***	0.128***	0.144***
	(0.034)	(0.049)	(0.039)	(0.050)
Week After, Treated Congruent	0.135***	0.192***	0.117***	0.160***
	(0.024)	(0.035)	(0.027)	(0.035)
Initial, Control Congruent	0.154***	0.179**	0.143**	0.133*
	(0.054)	(0.071)	(0.060)	(0.074)
Week After, Control Congruent	0.156***	0.100	0.136**	0.137*
	(0.050)	(0.069)	(0.057)	(0.071)
Immediate, Treated	-0.005	-0.015	-0.009	-0.009
	(0.019)	(0.027)	(0.019)	(0.026)
Week After, Treated	-0.018	-0.011	-0.016	0.001
	(0.013)	(0.022)	(0.014)	(0.022)
Initial, Control	0.013	-0.030	0.010	-0.015
	(0.049)	(0.058)	(0.049)	(0.058)
Week After, Control	0.127***	0.066	0.127***	0.040
	(0.048)	(0.057)	(0.048)	(0.056)
Mean	-0.00	0.00	-0.00	0.00
R2	0.44	0.18	0.44	0.18
Clusters	1,816	1,757	1,816	1,757
Obs.	8,642	7,806	8,642	7,806

Note: Outcome variable in columns (1) is a composite variable that aggregates the three different measures of beliefs elicited, it is standardized with mean zero and standard deviation one at every stage. Outcome variable in column (2) is the negative of the standardized WTA at every stage. The omitted category is initial treated. All specifications control for $\boldsymbol{Y}$: relative and absolute performance controls, and an indicator variable for the difficulty level of the first round tests; and $\boldsymbol{X}$: family income, indicators for each parent attending college, a nonwhite indicator, ACT scores, high school rank, indicator for attending high school in the U.S., honors student indicator, school year (freshman, sophomore, junior or senior), a measure of risk aversion, an indicator for taking the math quiz first, and immediate group indicator. In columns 1 and 2 congruent is equal to 1 if the domain is congruent with the the individuals gender (i.e. it is one for males when the domain is math, and one for females when the domain is verbal). In columns 3 and 4 congruent is equal to one if the participant believes that their gender has an advantage in that domain, zero otherwise. Errors clustered at individual level. Standard errors reported in parentheses. *Significant at 10%, **5%, ***1%.

Table D.16: Effect of Feedback on Expected Payoffs

	Initial	Immediately After	Week After
	(1)	(2)	(3)
Bad News$_D$	-0.719	-0.426	-0.849
	(0.560)	(0.833)	(0.543)
Bad News$_D$*Female	-0.003	-0.148	-0.065
	(0.267)	(0.353)	(0.240)
Good News$_D$*Female	-0.868*	-0.278	-1.061**
	(0.475)	(0.725)	(0.479)
Bad News$_D$*Congruent	0.130	-0.177	-0.018
	(0.212)	(0.289)	(0.193)
Good News$_D$*Congruent	-0.134	-0.104	-0.185
	(0.435)	(0.701)	(0.445)
Mean	5.84	5.86	5.81
R2	0.16	0.16	0.17
Clusters	1,453	689	1,453
Obs.	2,906	1,378	2,906

Note: Outcome variable is the expected payoff in dollars given participants decisions in each of the price lists at each stage (initial, immediate, week after). We randomly select a row from the price list math (verbal) from each of the three stages of the experiment (1,000 times for each stage), and calculate earnings based on the observed choice, then we average over the realized earnings. All specifications control for $\boldsymbol{Y}$: relative and absolute performance controls, and an indicator variable for the difficulty level of the first round tests; and $\boldsymbol{X}$: family income, indicators for each parent attending college, a nonwhite indicator, ACT scores, high school rank, indicator for attending high school in the U.S., honors student indicator, school year (freshman, sophomore, junior or senior), a measure of risk aversion, an indicator for taking the math quiz first and an indicator for being in Immediate group. D denotes an observation associated with a given domain. Errors clustered at individual level. Standard errors reported in parentheses. *Significant at 10%, **5%, ***1%.

Table D.17: Effect of Priors and Bayesian Posteriors on Expected Rank

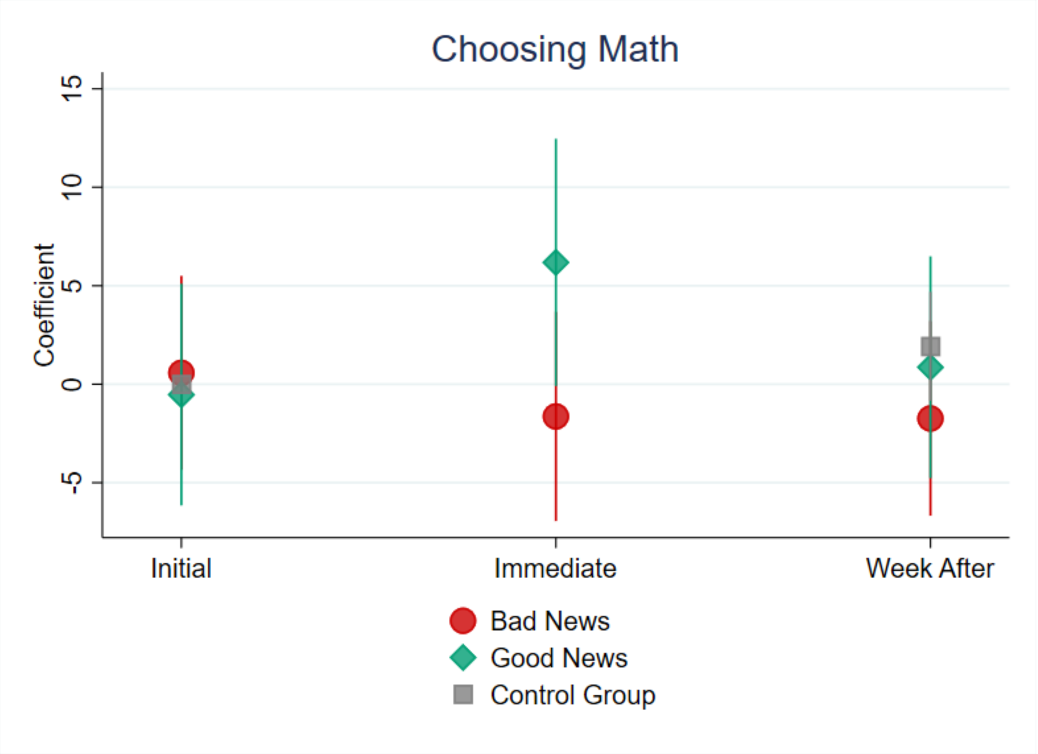
	Initial	Immediately After		Week After	
	(1)	(2)	(3)	(4)	(5)
Panel A: Prior Beliefs					
Bad News$_D$	0.059	-1.109***	-1.113***	-0.780***	-0.823***
	(0.110)	(0.160)	(0.101)	(0.103)	(0.065)
Bad News$_D$*Female	-0.716***	-0.601***	-0.197**	-0.683***	-0.156***
	(0.081)	(0.119)	(0.078)	(0.080)	(0.052)
Good News$_D$*Female	-0.588***	-0.339**	0.049	-0.278***	0.154***
	(0.092)	(0.132)	(0.070)	(0.088)	(0.052)
Bad News$_D$*Congruent	0.262***	0.161**	-0.022	0.191***	-0.001
	(0.059)	(0.077)	(0.047)	(0.051)	(0.035)
Good News$_D$*Congruent	0.230***	0.115	-0.080	0.154**	-0.014
	(0.076)	(0.103)	(0.060)	(0.071)	(0.044)
Prior Beliefs			✓		✓
R2	0.36	0.47	0.80	0.45	0.77
Panel B: Prior & Bayesian Posterior					
Bad News$_D$	0.059	-1.109***	-1.275***	-0.780***	-0.942***
	(0.110)	(0.160)	(0.147)	(0.103)	(0.104)
Bad News$_D$*Female	-0.716***	-0.601***	-0.200**	-0.683***	-0.162***
	(0.081)	(0.119)	(0.078)	(0.080)	(0.052)
Good News$_D$*Female	-0.588***	-0.339**	0.047	-0.278***	0.155***
	(0.092)	(0.132)	(0.070)	(0.088)	(0.052)
Bad News$_D$*Congruent	0.262***	0.161**	-0.024	0.191***	-0.002
	(0.059)	(0.077)	(0.047)	(0.051)	(0.035)
Prior Beliefs			✓		✓
Bayesian Posterior			✓		✓
R2	0.36	0.47	0.80	0.45	0.77
Mean	5.78	5.48	5.48	5.52	5.52
Clusters	1,453	689	689	1,453	1,453
Obs.	2,906	1,378	1,378	2,906	2,906

Note: Outcome variable is expected rank. It is regressed on an indicator for bad news, and interactions of good and bad news with female and congruent (i.e. omitted category is males who receive good news in gender-incongruent domains). Panel A controls for prior beliefs and Panel B, additionally, controls for Bayesina posterior. All specifications control for $\boldsymbol{Y}$: relative and absolute performance controls, and an indicator variable for the difficulty level of the first round tests; and $\boldsymbol{X}$: family income, indicators for each parent attending college, a nonwhite indicator, ACT scores, high school rank, indicator for attending high school in the U.S., honors student indicator, school year (freshman, sophomore, junior or senior), a measure of risk aversion, an indicator for taking the math quiz first and an indicator for being in the Immediate group. D denotes an observation associated with a given domain, and -D denotes the other domain. Errors clustered at individual level. Standard errors reported in parentheses. *Significant at 10%, **5%, ***1%.

APPENDIX E

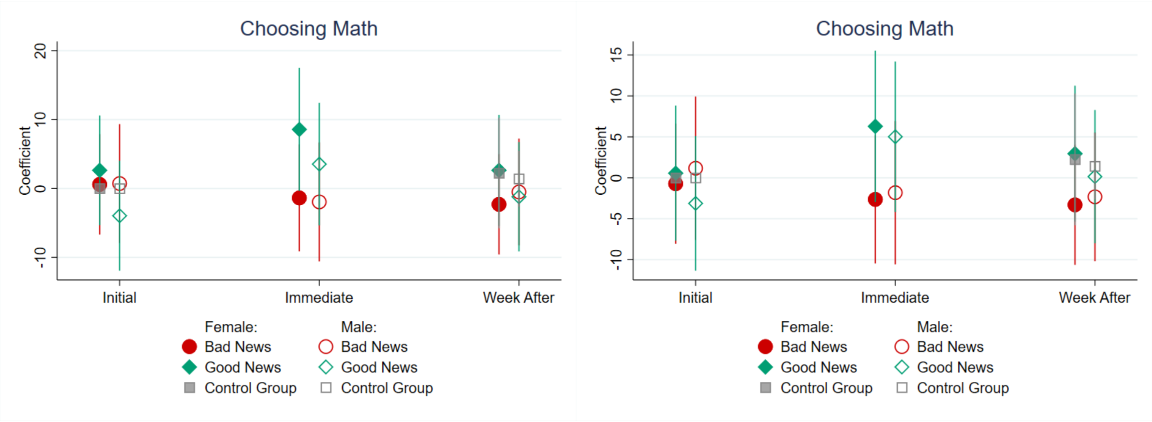
RESULTS FOR CHOOSING MATH, CHAPTER 3

Figure E.1: Levels over Time



Note: Outcome variable is an indicator variable equal to one when participant chooses to be compensated for math in Round 2. Markers represent the coefficient on good and bad news, and control group at different stages from a regression that pools all stages for all the participants. The outcome, indicated at the top of each panel, is regressed on indicator variables for initial good news, immediate good news, immediate bad news, week after good news, week after bad news and week after control group (i.e. the omitted category is initial control group), $\boldsymbol{Y}$: relative and absolute performance controls, an indicator variable for the difficulty level of the first round tests and all the controls in $\boldsymbol{X}$: gender, family income, indicators for each parent attending college, a nonwhite indicator, ACT scores, high school rank, indicator for attending high school in the U.S., honors student indicator, school year (freshman, sophomore, junior or senior), a measure of risk aversion, an indicator for taking the math quiz first, and an immediate group indicator. Errors clustered at individual level. The spikes represent 90% confidence intervals.

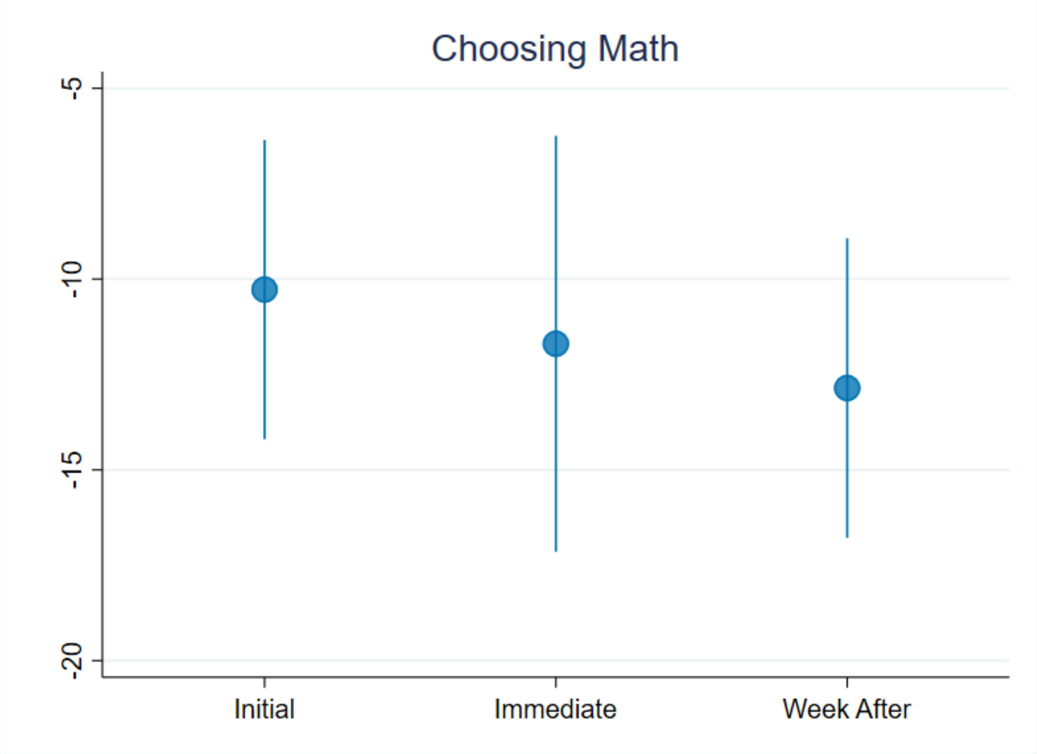
Figure E.2: Levels over Time by Gender


(a) All

(b) Accurate Subsample

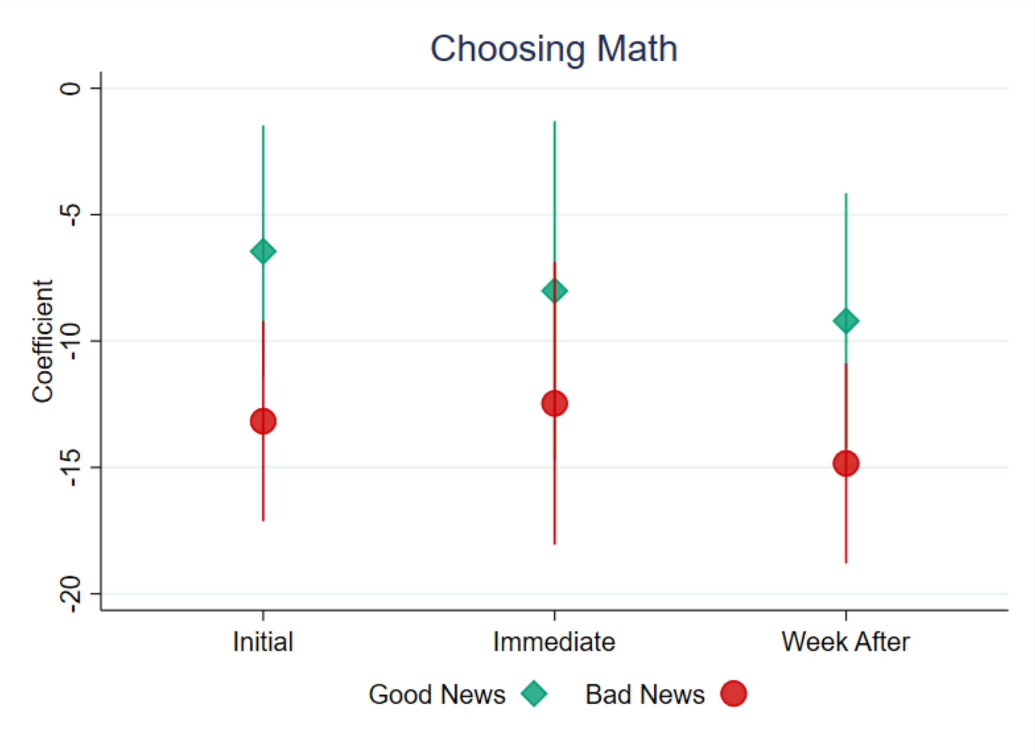
Note: Outcome variable is an indicator variable equal to one when participant chooses to be compensated for math in Round 2. Markers represent the coefficient on good and bad news, and control group at different stages for each gender from a regression that pools all stages for all the participants (only participants that accurately recalled feedback for Panel B). The outcome is regressed on indicator variables for initial good news male and female, immediate good news male and female, immediate bad news male and female, week after good news male and female, week after bad news male and female, initial control female and week after control group male and female (i.e. the omitted category is initial control male), $\boldsymbol{Y}$: relative and absolute performance controls, an indicator variable for the difficulty level of the first round tests and all the controls in $\boldsymbol{X}$: family income, indicators for each parent attending college, a nonwhite indicator, ACT scores, high school rank, indicator for attending high school in the U.S., honors student indicator, school year (freshman, sophomore, junior or senior), a measure of risk aversion, an indicator for taking the math quiz first, and an immediate group indicator. Errors clustered at individual level. The spikes represent 90% confidence intervals.

Figure E.3: Gaps over Time



Note: Outcome variable is an indicator variable equal to one when participant chooses to be compensated for math in Round 2. Markers represent the coefficient on the female indicator interacted with good and bad news at every stage from a regression that pools all stages for all the participants. The outcome is regressed on indicators for immediate treated good news, immediate treated bad news, week after treated good news, week after treated bad news, initial control, week after control (i.e. omitted category initial treated good news) and female interacted with those: initial treated good news female, initial treated bad news female, immediate treated good news female, immediate treated bad news female, week after treated good news female, week after treated bad news female, initial control female, week after control female, $\boldsymbol{Y}$: relative and absolute performance controls, an indicator variable for the difficulty level of the first round tests and all the controls in $\boldsymbol{X}$: family income, indicators for each parent attending college, a nonwhite indicator, ACT scores, high school rank, indicator for attending high school in the U.S., honors student indicator, school year (freshman, sophomore, junior or senior), a measure of risk aversion, an indicator for taking the math quiz first, and an indicator for being in the Immediate group. Errors clustered at individual level. The spikes represent 90% confidence intervals.

Figure E.4: Gaps over Time by Type of News



Note: Outcome variable is an indicator variable equal to one when participant chooses to be compensated for math in Round 2. Markers represent the coefficient on the female indicator interacted with good and bad news at every stage from a regression that pools all stages for all the participants. The outcome is regressed on indicators for immediate treated good news, immediate treated bad news, week after treated good news, week after treated bad news, initial control, week after control (i.e. omitted category initial treated good news) and female interacted with those: initial treated good news female, initial treated bad news female, immediate treated good news female, immediate treated bad news female, week after treated good news female, week after treated bad news female, initial control female, week after control female, $\boldsymbol{Y}$: relative and absolute performance controls, an indicator variable for the difficulty level of the first round tests and all the controls in $\boldsymbol{X}$: family income, indicators for each parent attending college, a nonwhite indicator, ACT scores, high school rank, indicator for attending high school in the U.S., honors student indicator, school year (freshman, sophomore, junior or senior), a measure of risk aversion, an indicator for taking the math quiz first, and an indicator for being in the Immediate group. Errors clustered at individual level. The spikes represent 90% confidence intervals.

Table E.1: Effect of Priors in Choosing Math

	Initial	Immediately After		Week After	
	(1)	(2)	(3)	(4)	(5)
Bad News$_{Math}$	5.913	-4.727	-10.312***	-2.252	-6.268*
	(4.523)	(6.387)	(3.680)	(4.431)	(3.371)
Bad News$_{Math}$*Female	-13.271***	-13.300***	-6.484**	-14.745***	-5.732**
	(3.135)	(4.424)	(2.634)	(3.129)	(2.353)
Good News$_{Math}$*Female	-6.567*	-7.346	-3.279	-9.459**	-4.999*
	(3.923)	(5.327)	(3.416)	(3.967)	(2.796)
Prior Chose Math			✓		✓
Mean	45.905	47.170	47.170	44.873	44.873
R2	0.243	0.309	0.739	0.243	0.593
Obs.	1,453	689	689	1,453	1,453

Note: Outcome variable is an indicator variable equal to one when participant chooses to be compensated for math in Round 2. Outcome is regressed on an indicator for bad news, and interactions of good and bad news with female (i.e. omitted category is males who receive good news). All specifications control for $\boldsymbol{Y}$: relative and absolute performance controls, and an indicator variable for the difficulty level of the first round tests; and $\boldsymbol{X}$: family income, indicators for each parent attending college, a nonwhite indicator, ACT scores, high school rank, indicator for attending high school in the U.S., honors student indicator, school year (freshman, sophomore, junior or senior), a measure of risk aversion, an indicator for taking the math quiz first and an indicator for being in Immediate group. Standard errors reported in parentheses. *Significant at 10%, **5%, ***1%.

Table E.2: Effect of Type of News on Choosing Math over Time

	(1)	(2)
Initial, Good News$_{Math}$	-0.529	-8.276**
	(3.421)	(3.783)
Initial, Bad News$_{Math}$	0.577	-8.280**
	(2.991)	(3.491)
Immediate, Good News$_{Math}$	6.185	-1.758
	(3.815)	(4.171)
Immediate, Bad News$_{Math}$	-1.637	-10.391***
	(3.227)	(3.636)
Week After, Good News$_{Math}$	0.860	-6.887*
	(3.421)	(3.761)
Week After, Bad News$_{Math}$	-1.741	-10.598***
	(3.000)	(3.486)
Week After, Control Group$_{Math}$	1.928	1.928
	(1.679)	(1.679)
Bad News$_{Verbal}$		13.255***
		(2.787)
Mean	46.22	46.22
R2	0.24	0.25
Clusters	1,816	1,816
Obs.	4,321	4,321

Note: Outcome variable is an indicator variable equal to one when participant chooses to be compensated for math in Round 2. The omitted category is initial control group. All specifications control for $\boldsymbol{Y}$: relative and absolute performance controls, and an indicator variable for the difficulty level of the first round tests; and $\boldsymbol{X}$: family income, indicators for each parent attending college, a nonwhite indicator, ACT scores, high school rank, indicator for attending high school in the U.S., honors student indicator, school year (freshman, sophomore, junior or senior), a measure of risk aversion, an indicator for taking the math quiz first, and immediate group indicator. Errors clustered at individual level. Standard errors reported in parentheses. *Significant at 10%, **5%, ***1%.

Table E.3: Effect of Type of News and Gender on Choosing Math over Time

	(1)
Initial, Good News$_{Math}$, Male	-3.954
	(4.847)
Initial, Bad News$_{Math}$, Male	0.722
	(4.679)
Initial, Good News$_{Math}$, Female	-10.403**
	(4.836)
Initial, Bad News$_{Math}$, Female	-12.448***
	(4.428)
Immediate, Good News$_{Math}$, Male	3.541
	(5.396)
Immediate, Bad News$_{Math}$, Male	-1.943
	(5.241)
Immediate, Good News$_{Math}$, Female	-4.472
	(5.430)
Immediate, Bad News$_{Math}$, Female	-14.411***
	(4.713)
Week After, Good News$_{Math}$, Male	-1.198
	(4.820)
Week After, Bad News$_{Math}$, Male	-0.501
	(4.706)
Week After, Good News$_{Math}$, Female	-10.403**
	(4.887)
Week After, Bad News$_{Math}$, Female	-15.342***
	(4.424)
Initial, Control Group$_{Math}$, Female	-13.045***
	(4.849)
Week After, Control Group$_{Math}$, Male	1.418
	(2.465)
Week After, Control Group$_{Math}$, Female	-10.793**
	(4.878)
Mean	46.22
R2	0.24
Clusters	1,816
Obs.	4,321

Note: Outcome variable is an indicator variable equal to one when participant chooses to be compensated for math in Round 2. The omitted category is initial control males. All specifications control for $\boldsymbol{Y}$: relative and absolute performance controls, and an indicator variable for the difficulty level of the first round tests; and $\boldsymbol{X}$: gender, family income, indicators for each parent attending college, a nonwhite indicator, ACT scores, high school rank, indicator for attending high school in the U.S., honors student indicator, school year (freshman, sophomore, junior or senior), a measure of risk aversion, an indicator for taking the math quiz first, and immediate group indicator. Errors clustered at individual level. Standard errors reported in parentheses. *Significant at 10%, **5%, ***1%.

Table E.4: Male-Female Gap over Time on Choosing Math Decision

	(1)
Initial, Treated Female	-10.275***
	(2.383)
Immediate, Treated Female	-11.696***
	(3.313)
Week After, Treated Female	-12.856***
	(2.386)
Immediate, Treated	1.715
	(3.203)
Week After, Treated	0.516
	(2.572)
Initial, Control Female	-12.751***
	(4.781)
Week After, Control Female	-11.917**
	(4.814)
Initial, Control	1.259
	(4.207)
Week After, Control	2.677
	(4.212)
Mean	46.22
R2	0.24
Obs.	4,321

Note: Outcome variable is an indicator variable equal to one when participant chooses to be compensated for math in Round 2. The omitted category is initial treated. All specifications control for $\boldsymbol{Y}$: relative and absolute performance controls, and an indicator variable for the difficulty level of the first round tests; and $\boldsymbol{X}$: family income, indicators for each parent attending college, a nonwhite indicator, ACT scores, high school rank, indicator for attending high school in the U.S., honors student indicator, school year (freshman, sophomore, junior or senior), a measure of risk aversion, an indicator for taking the math quiz first, and immediate group indicator. Errors clustered at individual level. Standard errors reported in parentheses. *Significant at 10%, **5%, ***1%.