A (Dynamic) Investigation of Stereotypes, Belief-Updating, and Behavior*

Katherine Coffman[†]
Harvard Business School

Maria Paola Ugalde Araya[‡]
Arizona State University

Basit Zafar[§]
University of Michigan

June 6, 2022

Abstract

Many decisions – such as what educational or career path to pursue – are dynamic in nature, with individuals receiving feedback at one point in time and making decisions later. Using a controlled experiment, with two sessions one week apart, we analyze the dynamic effects of feedback on beliefs about own performance and decision-making across two different domains (verbal skills and math). We 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. Despite this, we see substantial persistence of gender gaps over time. This is particularly true among the set of individuals who receive negative feedback. We find that, holding fixed performance and decisions before feedback, women update their beliefs and choices more negatively than men do after bad news. Our results highlight the challenges involved in overcoming gender gaps in dynamic settings.

*We thank Manuela Collis for excellent research assistance. We would also like to thank Esteban Aucejo, Ghazala Azmat, Lucas Coffman, Johanna Mollerstrom, Florian Zimmermann, and conference and seminar participants at Arizona State University, the CEBI Workshop on Subjective Beliefs in Macroeconomics and Household Finance, Stockholm School of Economics, University of Michigan, University of Toronto, the BRIQ Beliefs Workshop, the 2021 ESA North America Meeting, the University of Bonn Institute of Finance and Statistics, and the 2020 Workshop on Subjective Expectations for numerous helpful comments and suggestions. All errors that remain are ours.

[†]Harvard Business School, kcoffman@hbs.edu

[‡]Department of Economics, Arizona State University, mugaldea@asu.edu

[§]Department of Economics, University of Michigan, & NBER, basitak@gmail.com

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 paper, 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

[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.

individuals update their beliefs and 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. 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 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

[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.

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.

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 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.

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 the role of information in *static* settings. But, for many real-world

[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.

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

applications, belief formation, feedback, and choices are dynamic in nature. 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 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 (2019) 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 Ell, 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 Study Design and Administration

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 1 shows an overview diagram of the experimental design described in this section.

3.1.1 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 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,

[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.

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 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 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.

[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.

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 A1), 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 B, we present corresponding results for choosing math in the choice between the two piece-rate schemes.[7]

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

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

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

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 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.

3.1.2 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 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

[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.

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]

3.2 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 Session 1. A total of 2046 students completed

[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 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 A1). 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 A2). Panel A of Table 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 A2). 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.

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.

4.1 Descriptive Statistics

We first summarize the baseline data in Panel B of Table 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

[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.

Figure A2. 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 4, hold when we exclude these individuals at extreme ranks (see Appendix Table A3).

4.1.1 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 6.4 for women (versus an actual performance of 5.7). Panel A of Figure A3 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 A3 shows that the distributions do not differ significantly by gender (p-value of a K-S test=0.417). Panel B of Table 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 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 A3 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 A3 is 0.004, and in panel D is 0.01.[12]

Panel B of Table 1 shows that the mean subjective probability of ranking in the 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.

[12]Figure A4 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).

Initial choices show similar patterns by gender and domain. Table 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 A5 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]

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

4.1.2 The Provision of Feedback

Table A6 reports the percentages of participants that receive each possible feedback combination, separately by gender. Throughout this paper, 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 A7 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

[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 A4.

[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 of choice.

should revise her belief upward to just 8.29 after seeing good news, but downward to 7.00 after seeing bad news.

As a first pass, Table 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.

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.

4.2.1 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 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]

[15] Table A8 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.

[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 4, we omit the control group to focus on whether there is differential updating by gender, timing, and type of news.

Formally, our model is:

$$\begin{aligned} O_{iDt} =& \beta_0 + \beta_1 \textit{Initial Good News}_{iDt} + \beta_2 \textit{Immediate Good News}_{iDt} + \beta_3 \textit{Week After Good News}_{iDt} \\ &+ \beta_4 \textit{Initial Bad News}_{iDt} + \beta_5 \textit{Immediate Bad News}_{iDt} + \beta_6 \textit{Week After Bad News}_{iDt} \\ &+ \beta_7 \textit{Week After Control}_{iDt} + \mathbf{Y}_i + \mathbf{X}_i + \epsilon_{iDt}, \end{aligned} \tag{1}$$

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]

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

[19] Note that 79% of individuals make monotonic choices in every price list they see. In Appendix Table A9, we show that our results are quite similar even when we restrict attention to only those participants who are always monotonic.

[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 B. 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 A10 for the regression estimates that produces this figure. 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 A10 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 paper look quite similar. Results available upon request.

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 A8, 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] 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 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.

[24] This can be observed by subtracting the coefficient on Initial Good News from the coefficient 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, $p=0.67$.

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.

4.2.2 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 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

[27] In columns 3 and 4 of Table A11, 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 A12 for the regression estimates that produces this figure.

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 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 4. First, we point out that the positive trend in the control group that we observed in Figure 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 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 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)

[29] Table A11 shows the estimates that produces this figure.

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 A13, 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.

4.2.3 The Role of Recall

Figure 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 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, we regress a dummy for accurately recalling the feedback 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 A5 and A6, we show that the patterns we documented in Figures 4 and 5 do not

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

[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.

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.

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.

4.3.1 Plotting Gender Gaps Over Time

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

$$\begin{aligned} O_{iDt} =& \beta_0 + \beta_1 Immediate_{iDt} + \beta_2 Week\ After_{iDt} + \beta_3 Female_i \times Initial_{iDt} + \beta_4 Female_i \times Immediate_{iDt} \\ &+ \beta_5 Female_i \times Week\ After_{iDt} + \beta_6 Initial\ Control\ Group_{iDt} + \beta_7 Week\ After\ Control\ Group_{iDt} \\ &+ \beta_8 Female_i \times Initial\ Control\ Group_{iDt} + \beta_9 Female_i \times Week\ After\ Control\ Group_{iDt} \\ &+ \mathbf{Y}_i + \mathbf{X}_i + \epsilon_{iDt}, \end{aligned} \tag{2}$$

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 7.[32] Importantly, these are gaps controlling for performance. After completing the quizzes but prior 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,

[32] See Table A14 for the corresponding regression estimates.

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 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 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]

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

[34] See Table A15 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.

[35] In Table A16, we show that our intervention also has a minimal overall impact on gender gaps in payoffs. In addition, Figure A7 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.

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

$$\begin{aligned} O_{iDt} =& \beta_0 + \beta_1 \textit{Immediate Good News}_{iDt} + \beta_2 \textit{Week After Good News}_{iDt} + \beta_3 \textit{Female}_i \times \textit{Initial Good News}_{iDt} \\ &+ \beta_4 \textit{Female}_i \times \textit{Immediate Good News}_{iDt} + \beta_5 \textit{Female}_i \times \textit{Week After Good News}_{iDt} + \beta_6 \textit{Initial Bad News}_{iDt} \\ &+ \beta_7 \textit{Immediate Bad News}_{iDt} + \beta_8 \textit{Week After Bad News}_{iDt} + \beta_9 \textit{Female}_i \times \textit{Initial Bad News}_{iDt} \\ &+ \beta_{10} \textit{Female}_i \times \textit{Immediate Bad News}_{iDt} + \beta_{11} \textit{Female}_i \times \textit{Week After Bad News}_{iDt} \\ &+ \beta_{12} \textit{Initial Control Group}_{iDt} + \beta_{13} \textit{Week After Control Group}_{iDt} + \beta_{14} \textit{Female}_i \times \textit{Initial Control Group}_{iDt} \\ &+ \beta_{15} \textit{Female}_i \times \textit{Week After Control Group}_{iDt} + \mathbf{Y}_i + \mathbf{X}_i + \epsilon_{iDt}. \end{aligned} \tag{3}$$

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 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 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 9 and 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:

$$\begin{aligned} O_{iD} =& \beta_0 + \beta_1 \textit{Bad News}_{iD} + \beta_2 \textit{Bad News}_{iD} \times \textit{Female}_i + \beta_3 \textit{Good News}_{iD} \times \textit{Female}_i \\ &+ \beta_4 \textit{Bad News}_{iD} \times \textit{Congruent}_{iD} + \beta_5 \textit{Good News}_{iD} \times \textit{Congruent} + \beta_6 \textit{Prior } O_{iD} + \mathbf{Y}_i + \mathbf{X}_i + \epsilon_{iD}, \end{aligned} \tag{4}$$

where $D \in \{\text{Verbal, Math}\}$, O_{iD} is a measure of beliefs or -WTA for participant i in domain D, and *Prior* O_{iD} is the initial outcome (belief or -WTA) when the model is estimated using the immediate or week after decisions. $\textit{Congruent}_{iD}$ 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

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

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 4 presents the results for beliefs. Column (1) estimates the initial 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 9 and 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 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 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 A17, 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 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 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.

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 paper, 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.

References

(2018, August). Dynamics of the gender gap in high math achievement. Working Paper 24910, National Bureau of Economic Research.

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

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

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

Benjamin, D. J. (2019). Chapter 2 - errors in probabilistic reasoning and judgment biases. In B. D. Bernheim, S. DellaVigna, and D. Laibson (Eds.), *Handbook of Behavioral Economics - Foundations and Applications 2*, Volume 2 of *Handbook of Behavioral Economics: Applications and Foundations 1*, pp. 69–186. North-Holland.

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

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

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

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

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

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

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

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

Buser, T. and H. Yuan (2019). 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.

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

Cason, T. N., W. A. Masters, and R. M. Sheremeta (2010). Entry into winner-take-all and proportional-prize contests: An experimental study. *Journal of Public Economics 94*(9), 604–611.

Charness, G. and C. Dave (2017). Confirmation bias with motivated beliefs. *Games and Economic Behavior 104*(C), 1–23.

Chen, D. L., M. Schonger, and C. Wickens (2016). otree—an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance 9*, 88 – 97.

Chew, S. H., W. Huang, and X. Zhao (2020). Motivated False Memory. *Journal of Political Economy 128*(10), 3913–3939.

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

Coffman, K. B. (2014). Evidence on Self-Stereotyping and the Contribution of Ideas. *The Quarterly Journal of Economics 129*(4), 1625–1660.

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

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

Eil, D. and J. M. Rao (2011, May). The good news-bad news effect: Asymmetric processing of objective information about yourself. *American Economic Journal: Microeconomics 3*(2), 114–38.

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

Ertac, S. and B. Szentes (2011, February). 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.

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

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

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

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

Gotthard-Real, A. (2017). Desirability and information processing: An experimental study. *Economics Letters 152*(C), 96–99.

Grossman, Z. and D. Owens (2012). An unlucky feeling: Overconfidence and noisy feedback. *Journal of Economic Behavior & Organization 84*(2), 510–524.

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

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

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

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

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

Mezulis, A., L. Abramson, J. Hyde, and B. Hankin (2004). 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.

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

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

Niederle, M. and L. Vesterlund (2007). Do Women Shy Away From Competition? Do Men Compete Too Much? *The Quarterly Journal of Economics 122*(3), 1067–1101.

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

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

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

Reuben, E., M. Wiswall, and B. Zafar (2017). Preferences and biases in educational choices and labour market expectations: Shrinking the black box of gender. *The Economic Journal 127*(604), 2153–2186.

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

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

Schwardmann, P. and J. Weele (2019, 10). Deception and self-deception. *Nature Human Behaviour 3*.

Shastry, G. K., O. Shurchkov, and L. L. Xia (2020). Luck or skill: How women and men react to noisy feedback. *Journal of Behavioral and Experimental Economics 88*, 101592.

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

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

Zimmermann, F. (2020, February). The dynamics of motivated beliefs. *American Economic Review 110*(2), 337–61.

Figures and Tables

Figure 1: Experimental Design

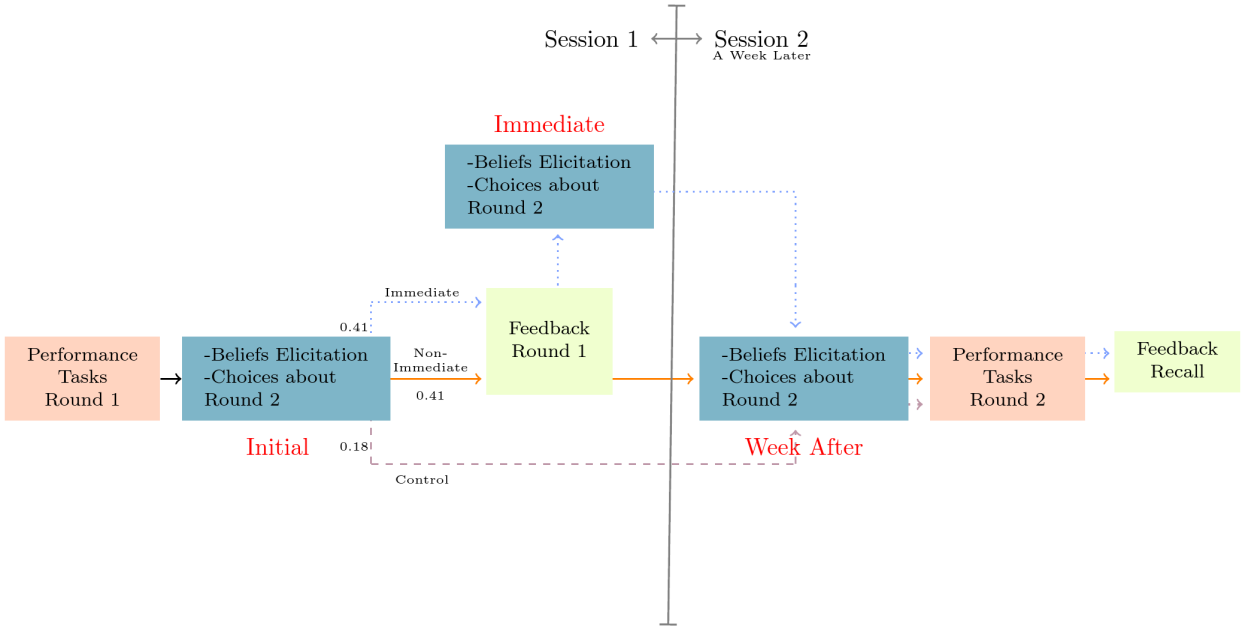
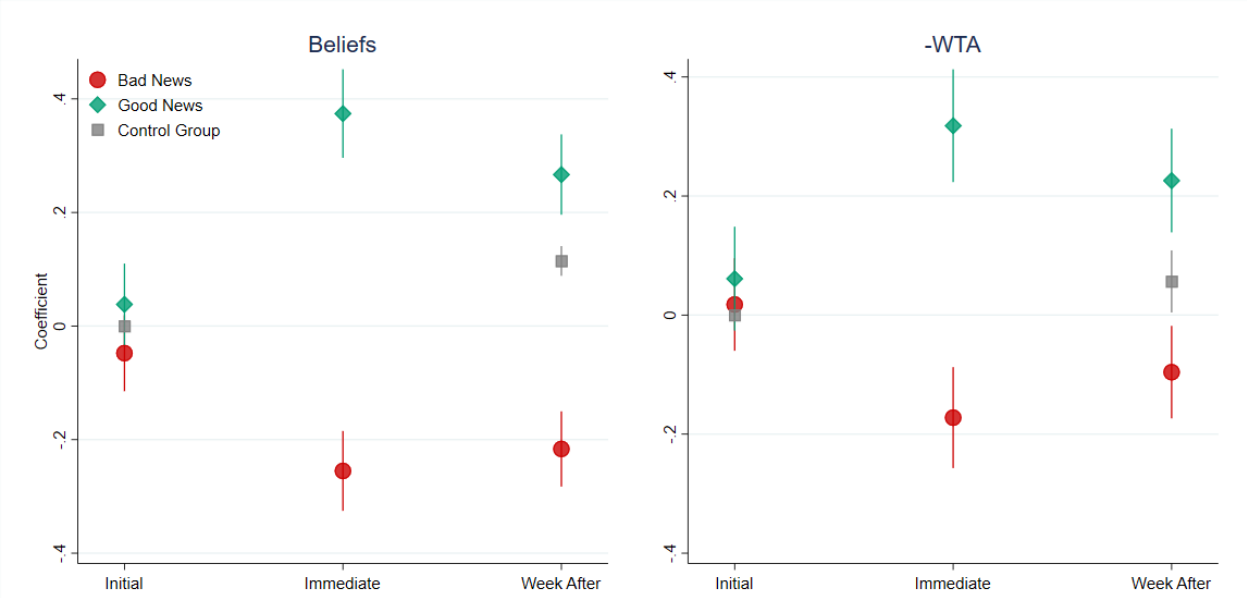


Figure 2: Price Lists for Round 2 Payments, Math

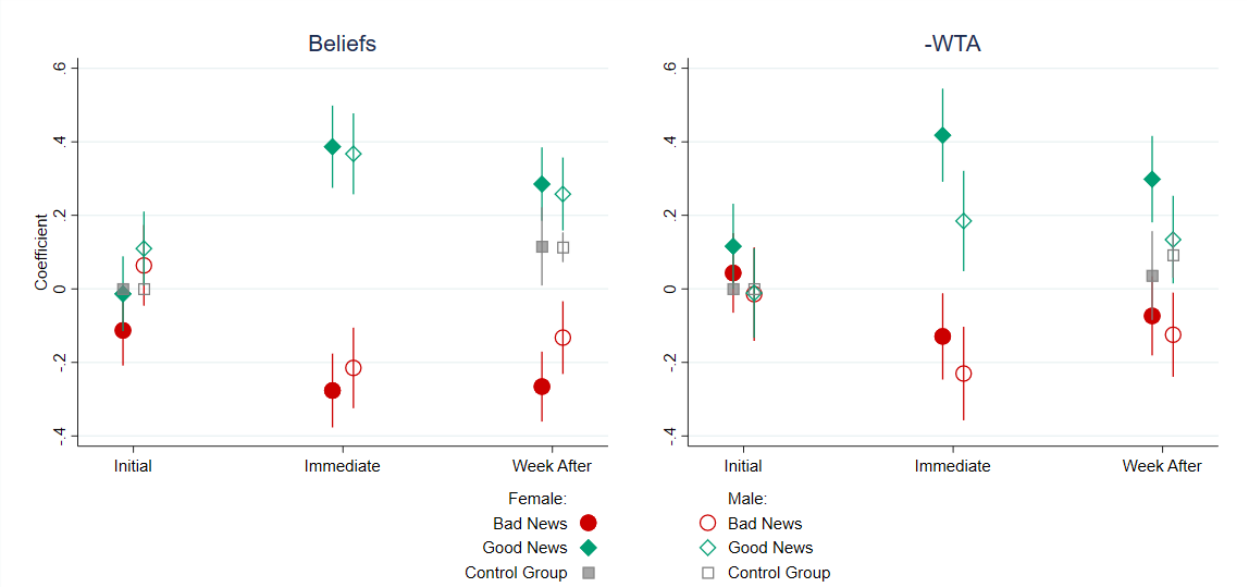
First Option	Second Option
Earn \$4 per problem solved correctly on the **Math** test in **Round 2** if in top-4; \$0 otherwise.	\$1 per problem solved correctly on the **Verbal** test in **Round 2**.
Earn \$3.50 per problem solved correctly on the **Math** test in **Round 2** if in top-4; \$0 otherwise.	\$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: Levels over Time



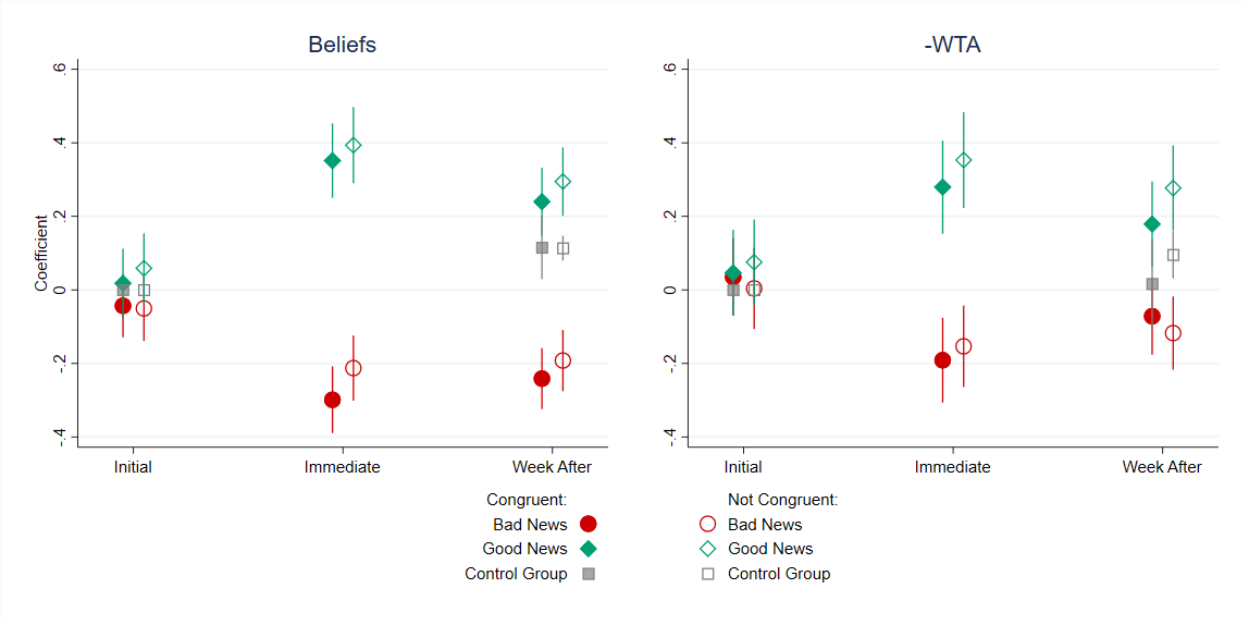
Note: Markers represent the coefficient on good and bad news, and control group at different stages from a regression that pools all stages for all the participants. The outcome, indicated at the top of each panel, is regressed on indicator variables for initial good news, immediate good news, immediate 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 standard normal distributed with mean zero and standard deviation one, per stage. -WTA is the negative of the WTA standardized per stage. The spikes represent 90% confidence intervals.

Figure 4: Levels over Time by Gender



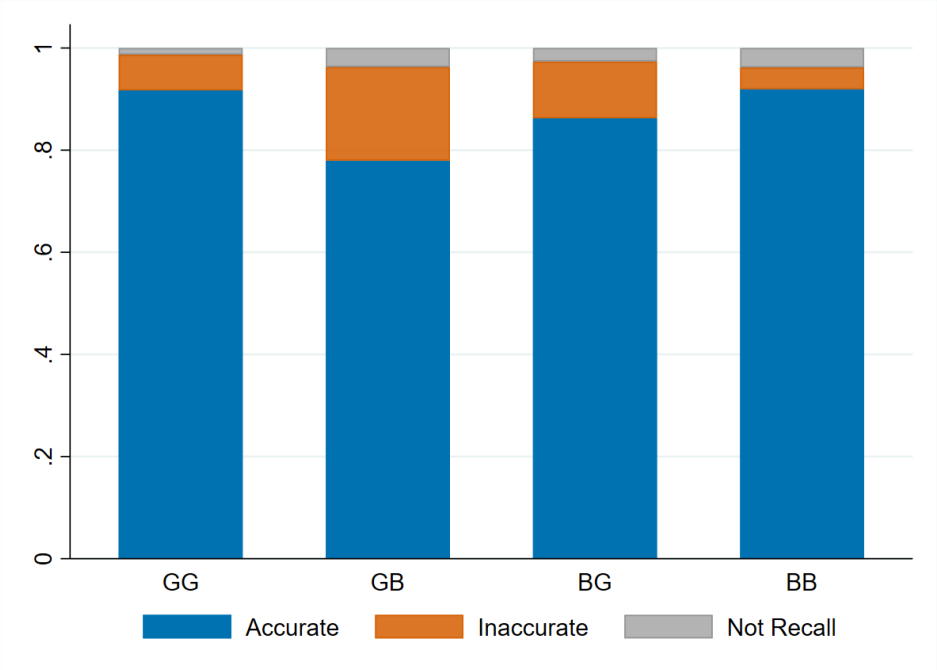
Note: Markers represent the coefficient on good and bad news, and control group at different stages for each gender from a regression that pools all stages for all the participants. The outcome, indicated at the top of each panel, is regressed on indicator variables for initial good news male and female, 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), $\boldsymbol{Y}$: 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. 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 standard normal distributed with mean zero and standard deviation one, per stage. -WTA is the negative of the WTA standardized per stage. The spikes represent 90% confidence intervals.

Figure 5: Levels over Time by Domain Congruence



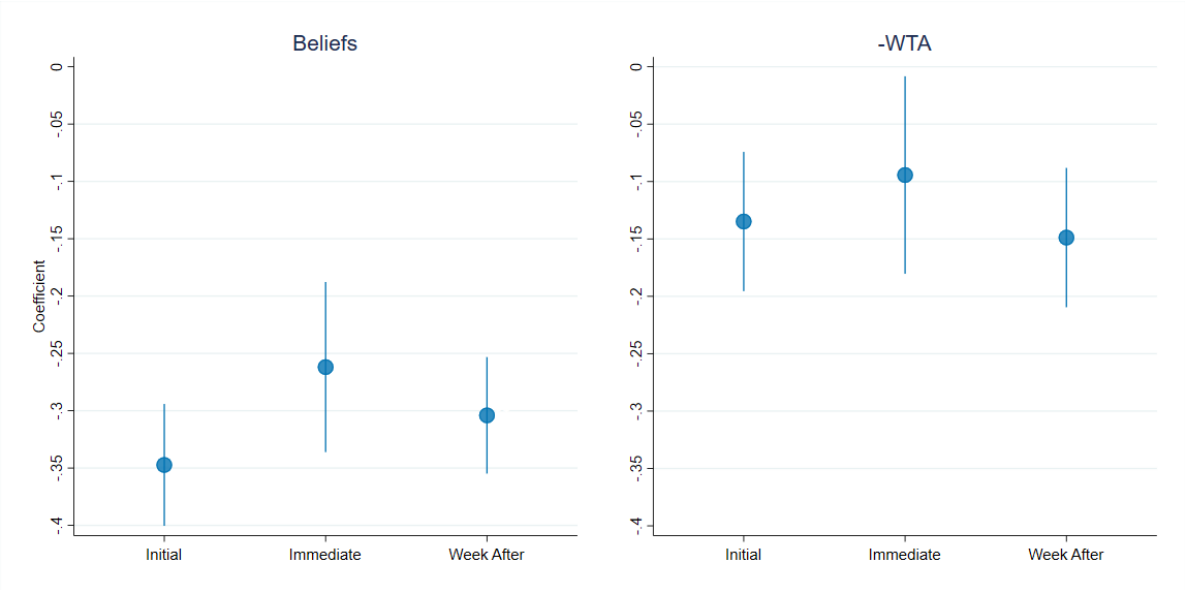
Note: Markers represent the coefficient on good and bad news, and control group at different stages for congruent and incongruent domains from a regression that pools all stages for all the participants. The outcome, indicated at the top of each panel, is regressed on indicator variables for initial 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 standard normal distributed with mean zero and standard deviation one, per stage. -WTA is the negative of the WTA standardized per stage. The spikes represent 90% confidence intervals.

Figure 6: Feedback Recall by Type of News



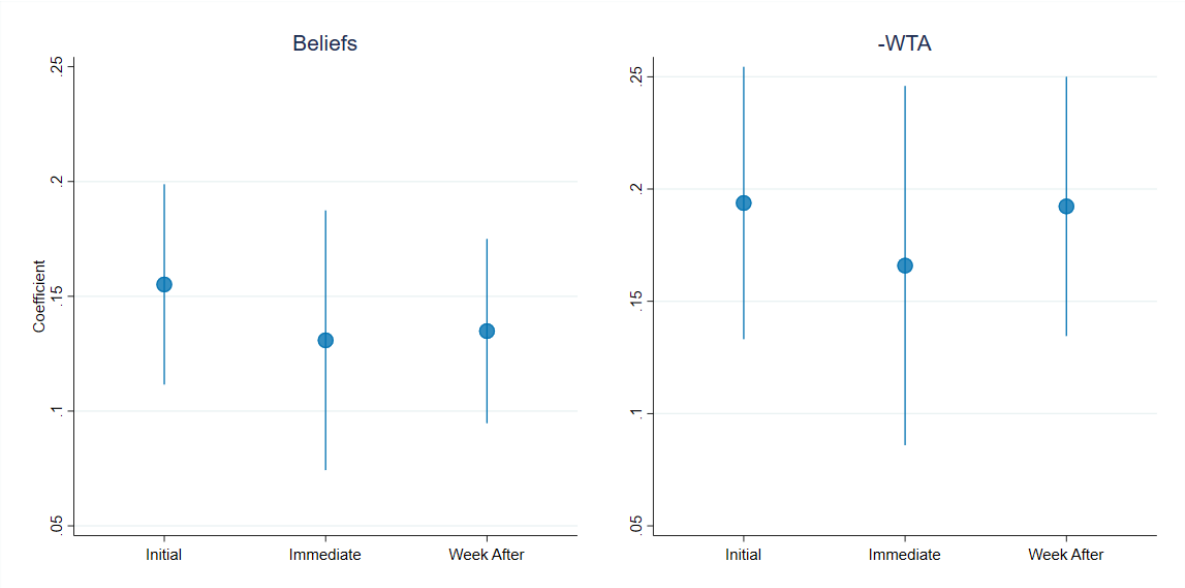
Note: GG: good news in both domains. BB: bad news in both domains. GB: good news in D, bad news in -D. BG: bad news in D, good news in -D. D denotes an observation associated with a given domain, and -D denotes the other domain. Standard errors reported in parentheses.

Figure 7: Male-Female Gap over Time



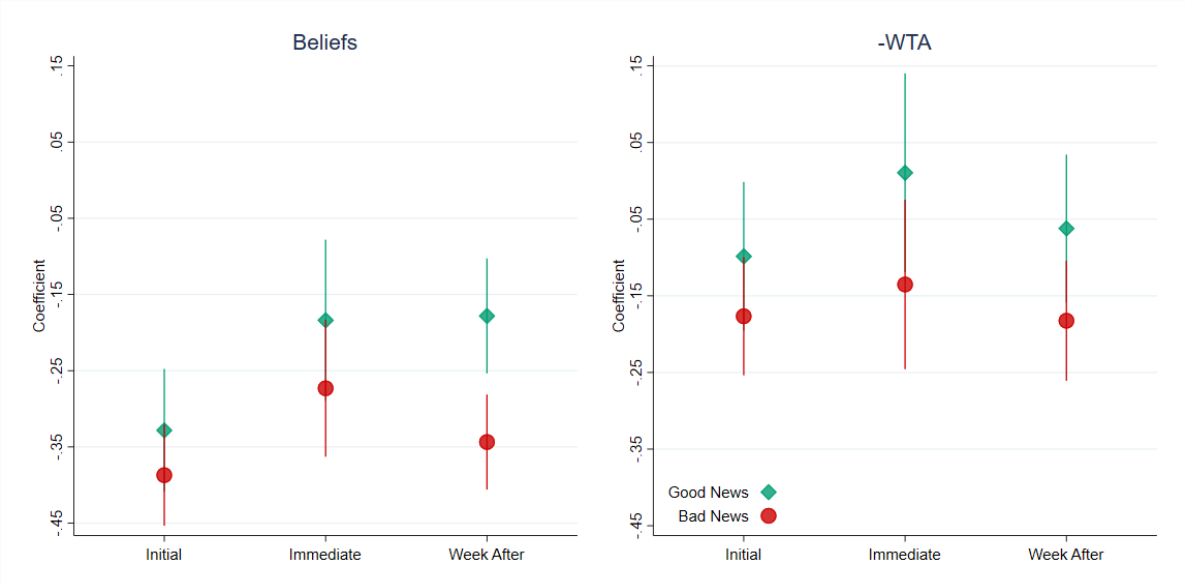
Note: Markers represent the coefficient on the female indicator interacted with initial, immediate and week after indicators (for females not in the control group) from a regression that pools all stages for all the participants. The outcome, indicated at the top of each panel, is regressed on indicators 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 standard normal distributed with mean zero and standard deviation one, per stage. -WTA is the negative of the WTA standardized per stage. The spikes represent 90% confidence intervals.

Figure 8: Stereotype Differences over Time



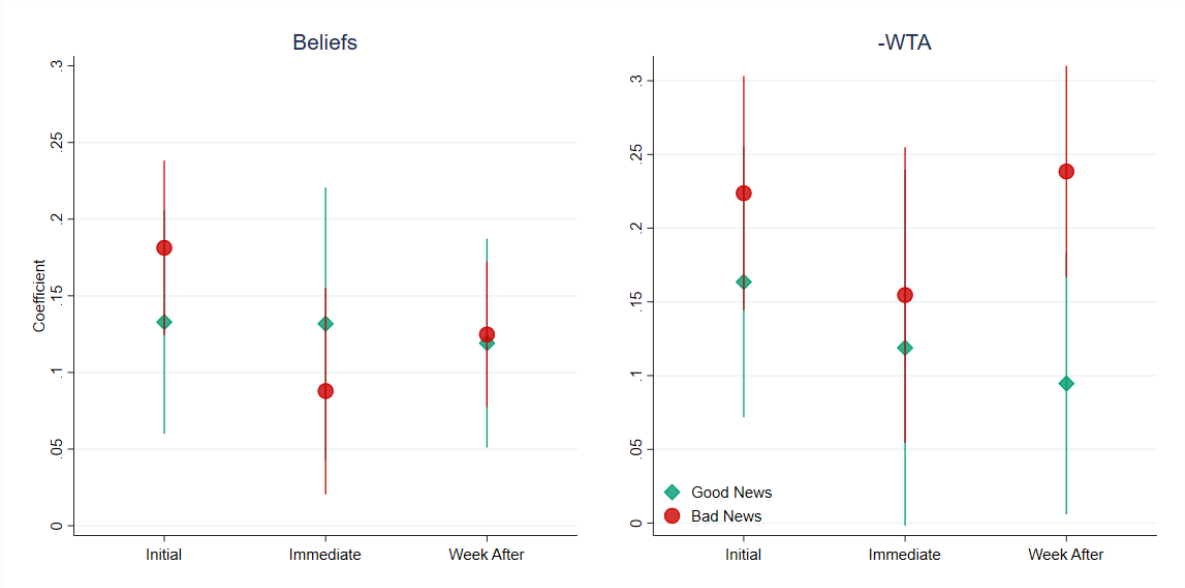
Note: Markers represent the coefficient on the domain congruence indicator interacted with initial, immediate and week after indicators (for participants not in the control group) from a regression that pools all stages for all the participants. The outcome, indicated at the top of each panel, is regressed 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 standard normal distributed with mean zero and standard deviation one, per stage. -WTA is the negative of the WTA standardized per stage. The spikes represent 90% confidence intervals.

Figure 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 standard normal distributed with mean zero and standard deviation one, per stage. -WTA is the negative of the WTA standardized per stage. The spikes represent 90% confidence intervals.

Figure 10: Stereotype Differences over Time by Type if 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 standard normal distributed with mean zero and standard deviation one, per stage. -WTA is the negative of the WTA standardized per stage. The spikes represent 90% confidence intervals.

Table 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 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: 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 $\boldsymbol{Y}$: relative and absolute performance controls, and an indicator variable for the difficulty level of the first round tests; time spent during Session 1, and $\boldsymbol{X}$: 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. Standard errors reported in parentheses. Errors clustered at individual level. *Significant at 10%, **5%, ***1%.

Table 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 standard normal distributed with mean zero and standard deviation one per 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%.

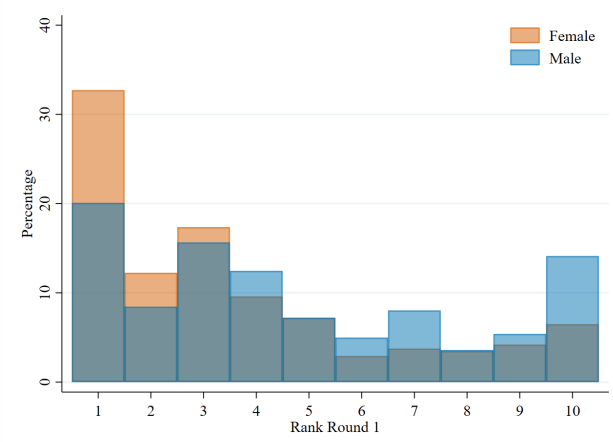
ONLINE APPENDIX

Figure A1: Price Lists for Round 2 Payments, Verbal

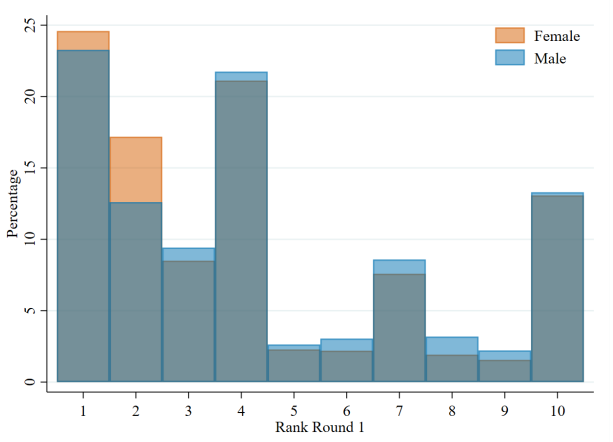
First Option	Second Option
$1 per problem solved correctly on the **Math** test in **Round 2**.	Earn $1.5 per problem solved correctly on the **Verbal** test in **Round 2** if in top-4; $0 otherwise.
$1 per problem solved correctly on the **Math** test in **Round 2**.	Earn $2 per problem solved correctly 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 A2: Round 1 Actual Rank Histograms

(a) Math

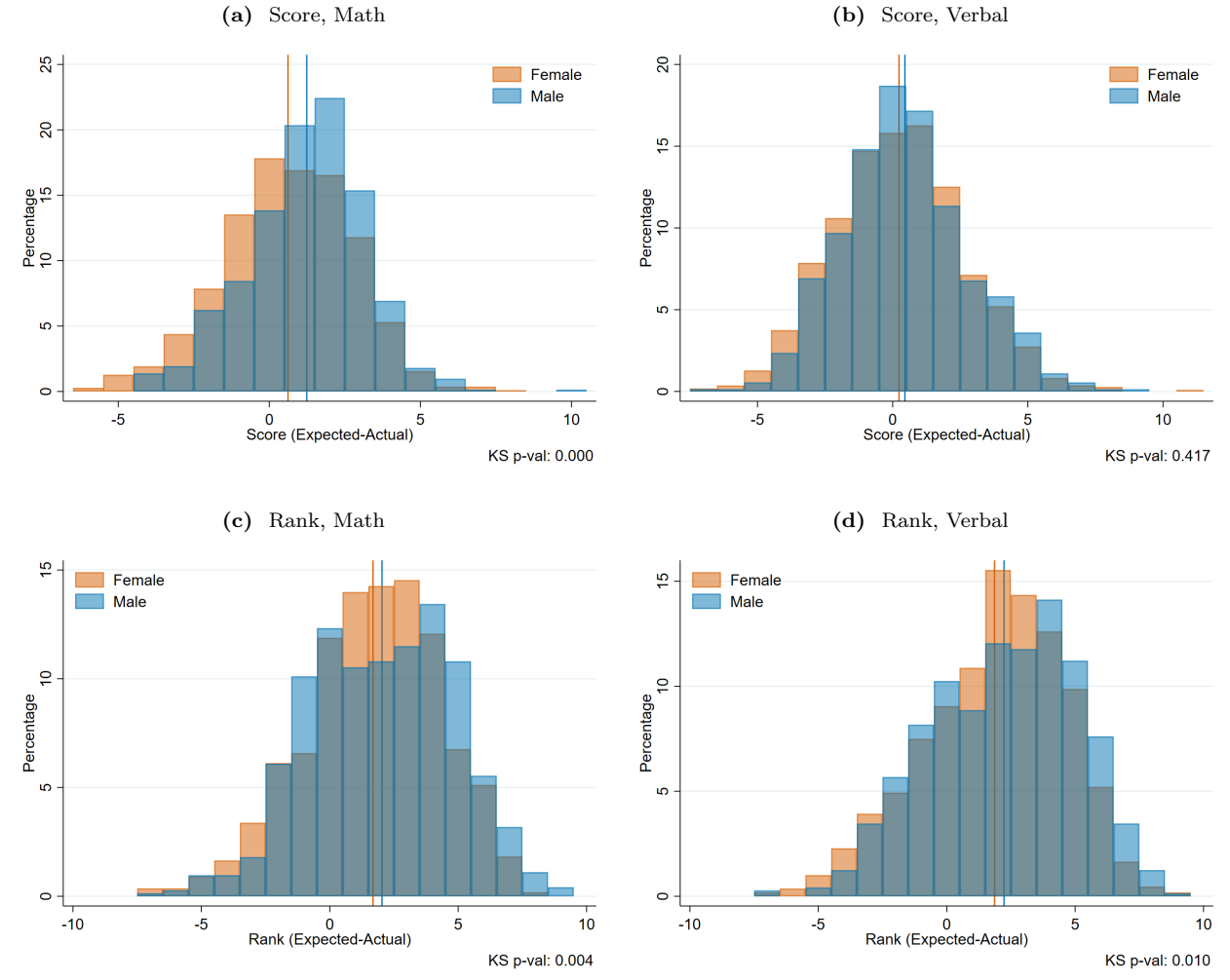


(b) Verbal



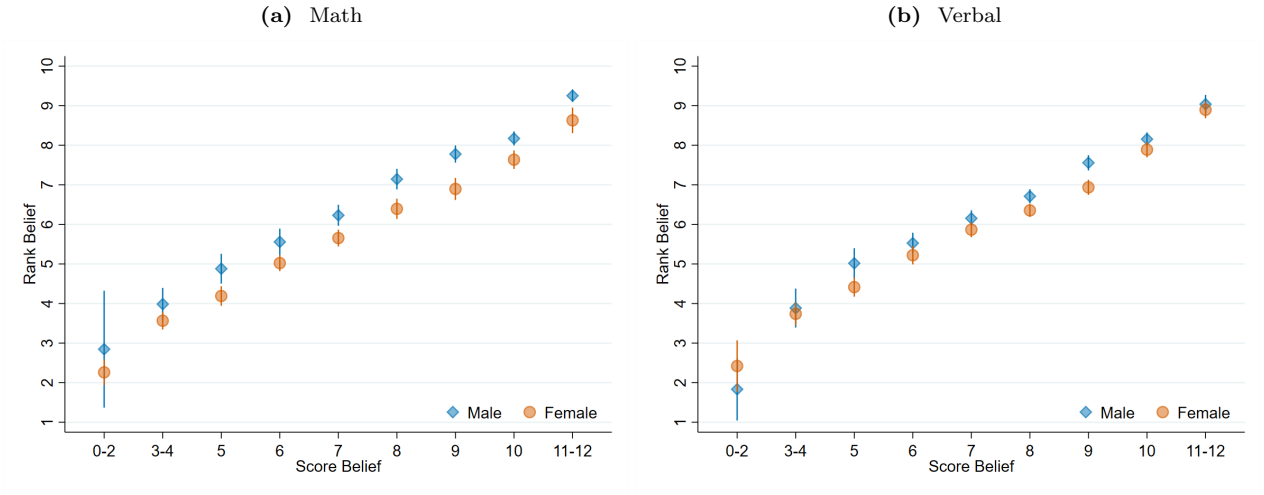
Note: Best ranking is ten, worst is one.

Figure A3: Relative and Absolute Overconfidence by Gender



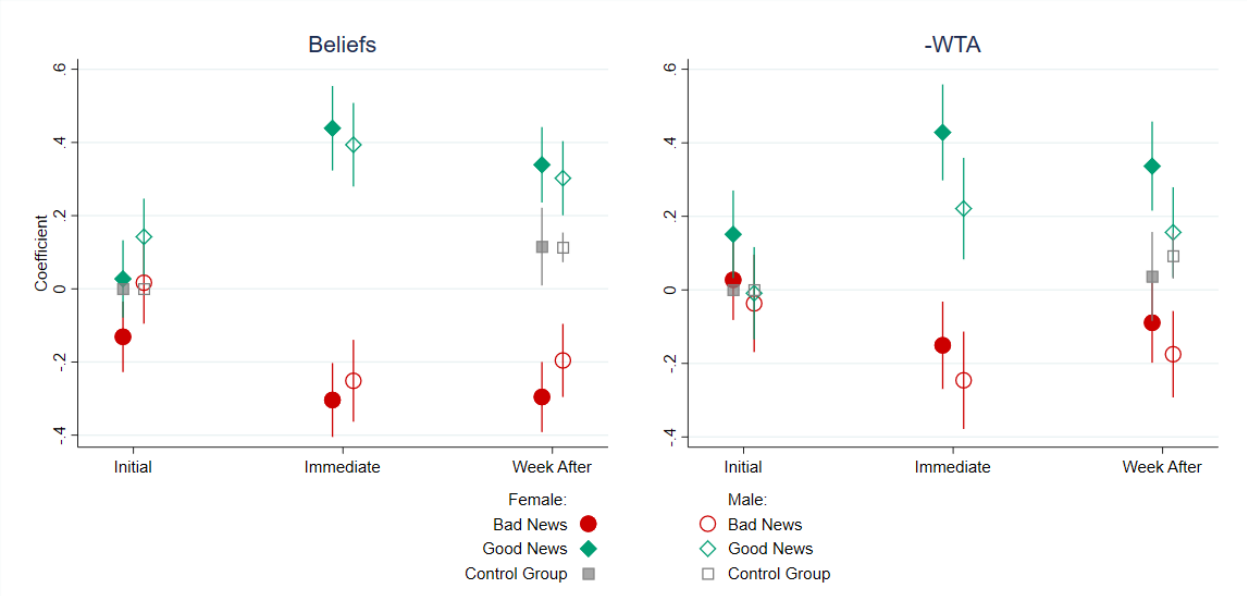
Note: Panel (a) and (b) are histograms of the difference between the initial expected number of correct answers and the actual number of correct answers in each Round 1 quiz, respectively. Panel (c) and (d) are histograms of the difference between the initial expected rank and the actual rank of participants 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 A4: Rank Belief by Score Belief



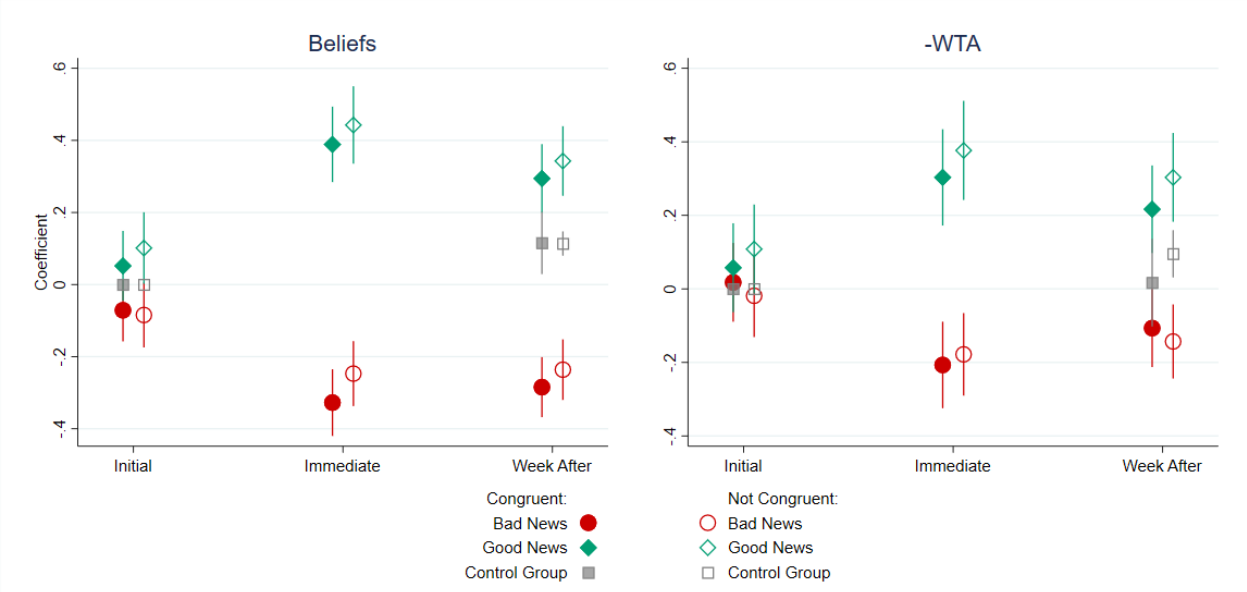
Note: Markers represent the mean of the rank beliefs by each score level on the x-axis. The spikes represent 95% confidence intervals.

Figure A5: Levels over Time by Gender, Accurate Subsample



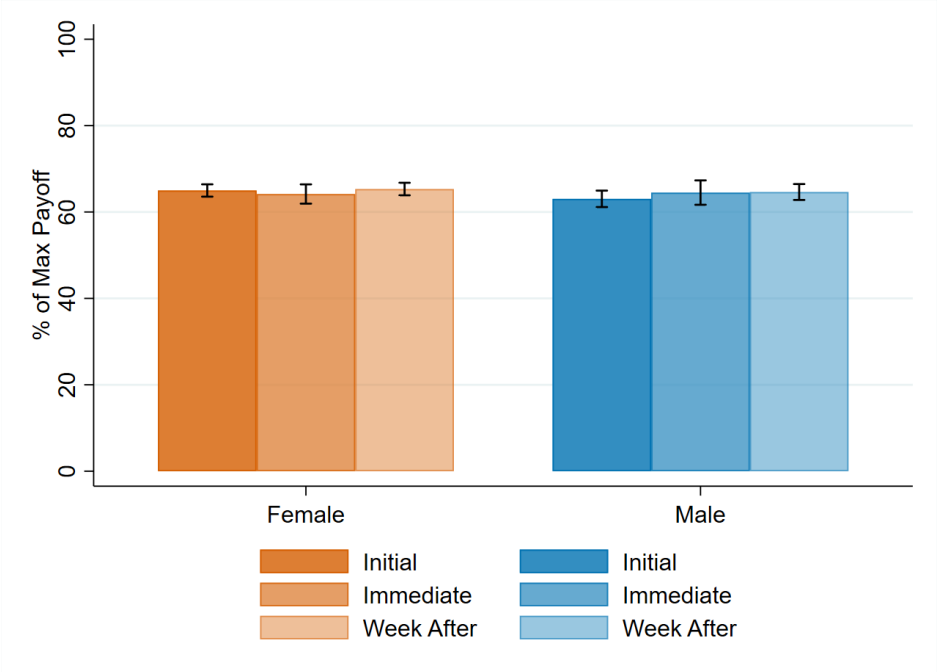
Note: Markers represent the coefficient on good and bad news, and control group at different stages for each gender from a regression that pools all stages for all the participants that accurately remember their feedback for that domain. The outcome, indicated at the top of each panel, is regressed on 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 standard normal distributed with mean zero and standard deviation one, per stage. -WTA is the negative of the WTA standardized per stage. The spikes represent 90% confidence intervals.

Figure A6: 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 standard normal distributed with mean zero and standard deviation one, per stage. -WTA is the negative of the WTA standardized per stage. The spikes represent 90% confidence intervals.

Figure A7: 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 A1: 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 varible 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 A2: 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]

Notes: 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 A3: 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 standard normal distributed with mean zero and standard deviation one per 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 A4: 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 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 A1, the price list for the competitive payment scheme in verbal vs \$1 math.

Table A5: 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 standard normal distributed with mean zero and standard deviation one. 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 A6: 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 A7: 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 A8: 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 A9: 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 standard normal distributed with mean zero and standard deviation one per 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 A10: 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$_{-D}$		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 standard normal distributed 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 A11: 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 standard normal distributed 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 A12: 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 standard normal distributed 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 A13: Posterior Expected Rank on Bayesian Update

	Immediately After Feedback								Week After							
	Math				Verbal				Math				Verbal			
	Female		Male		Female		Male		Female		Male		Female		Male	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
Bayes	0.874***	0.877***	0.882***	0.850***	0.772***	0.700***	0.841***	0.820***	0.846***	0.851***	0.840***	0.768***	0.770***	0.713***	0.818***	0.766***
	(0.030)	(0.036)	(0.040)	(0.050)	(0.038)	(0.046)	(0.043)	(0.052)	(0.023)	(0.027)	(0.026)	(0.031)	(0.025)	(0.030)	(0.028)	(0.034)
Good News$_D$	0.145	0.223	-0.203	-0.803	0.247*	-1.025**	-0.077	-0.513	-0.104	0.030	-0.345***	-1.837***	0.067	-1.001***	-0.278***	-1.396***
	(0.125)	(0.422)	(0.152)	(0.577)	(0.142)	(0.507)	(0.155)	(0.632)	(0.097)	(0.338)	(0.101)	(0.383)	(0.095)	(0.340)	(0.100)	(0.416)
Good News$_D$*Bayes		-0.013		0.090		0.206***		0.066		-0.022		0.220***		0.173***		0.166***
		(0.067)		(0.083)		(0.079)		(0.092)		(0.053)		(0.055)		(0.053)		(0.060)
Constant	0.584***	0.568***	0.748***	0.909***	1.050***	1.376***	0.997***	1.102***	0.893***	0.870***	1.072***	1.440***	1.245***	1.500***	1.190***	1.459***
	(0.138)	(0.161)	(0.214)	(0.261)	(0.184)	(0.221)	(0.228)	(0.271)	(0.102)	(0.116)	(0.142)	(0.167)	(0.119)	(0.142)	(0.152)	(0.179)
Mean	4.936	4.936	6.062	6.062	5.345	5.345	5.906	5.906	4.959	4.959	6.118	6.118	5.370	5.370	6.009	6.009
R2	0.774	0.774	0.748	0.749	0.668	0.673	0.698	0.699	0.720	0.720	0.731	0.738	0.672	0.676	0.692	0.696
N	411	411	278	278	411	411	278	278	872	872	581	581	872	872	581	581
Estimated Responsiveness to:																
Good News		0.864		0.940		0.906		0.886		0.829		0.988		0.886		0.932
Bad News		0.877		0.850		0.700		0.820		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 A14: 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 standard normal distributed 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. Errors clustered at individual level. Standard errors reported in parentheses. *Significant at 10%, **5%, ***1%.

Table A15: 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 standard normal distributed 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 A16: 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 A17: 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%.

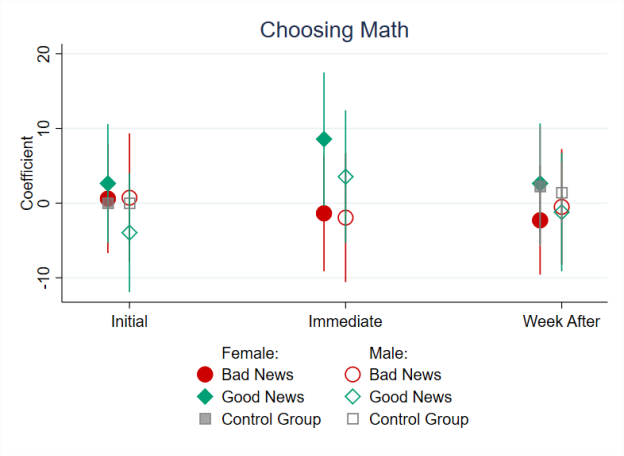
B Appendix: Choosing Math

Figure B1: Levels over Time

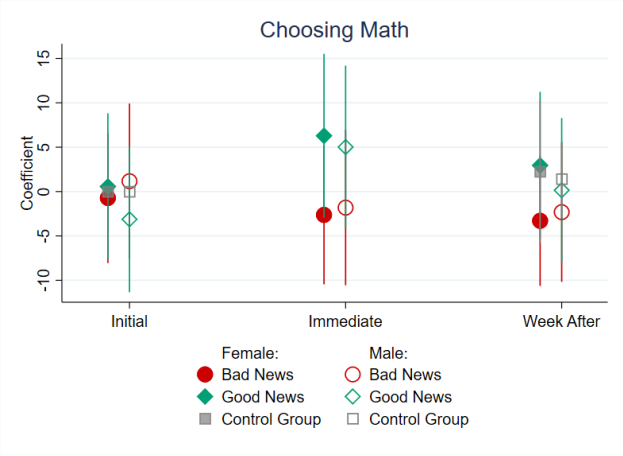
Note: Outcome variable is an indicator variable equal to one when participant chooses to be compensated for math in Round 2. Markers represent the coefficient on good and bad news, and control group at different stages from a regression that pools all stages for all the participants. The outcome, indicated 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 B2: Levels over Time by Gender

(a) All

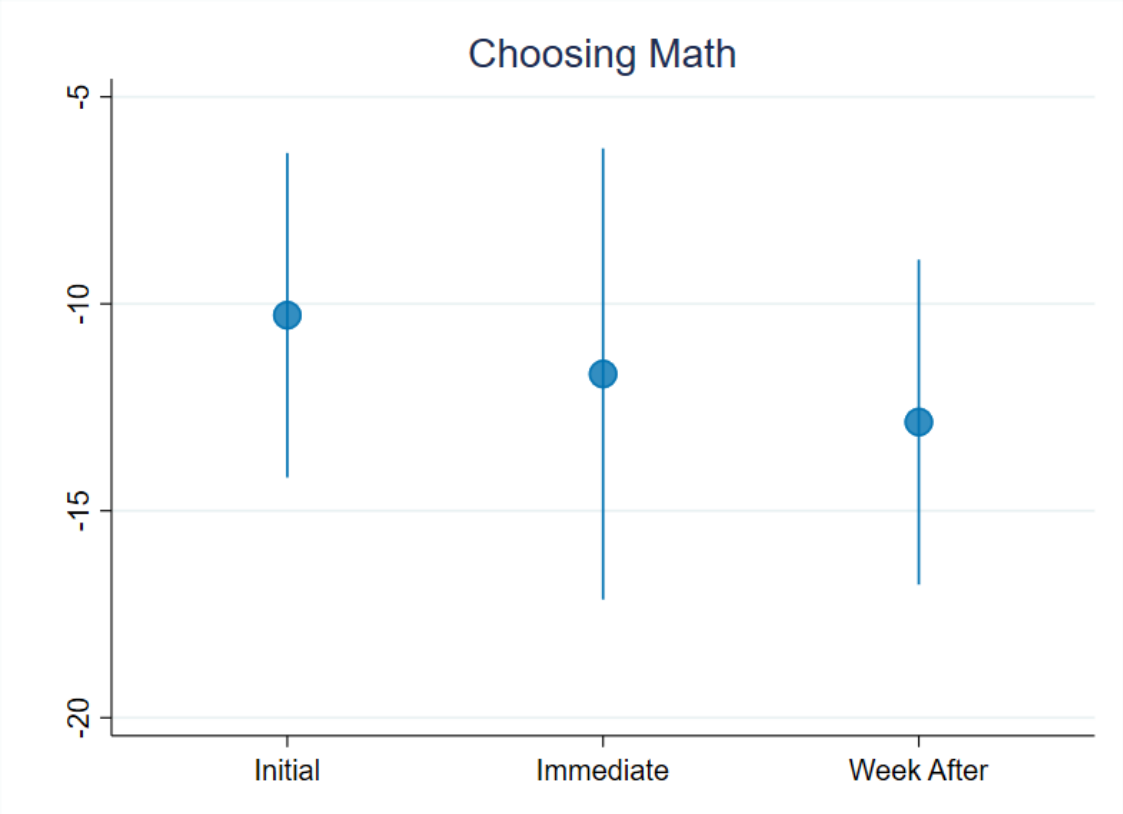


(b) Accurate Subsample



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 B3: 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 B4: 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 B1: 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 B2: 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 B3: 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 B4: 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%.