STEREOTYPES AND BELIEF UPDATING*

Katherine Coffman[ǂ]

Manuela R. Collis

Leena Kulkarni

September 2023

Abstract: We explore how feedback shapes, and perpetuates, gender gaps in self-assessments. Participants in our experiment take tests of their ability across different domains. We elicit their beliefs of their performance before and after feedback. We find that, even after the provision of highly informative feedback, gender stereotypes influence posterior beliefs, beyond what a Bayesian model would predict. This is primarily because both men and women update their beliefs more positively in response to good news when it arrives in a more gender congruent domain (i.e., more male-typed domains for men, more female-typed domains for women), fueling persistence in gender gaps.

JEL Codes: C91, D83

*This work was funded by the National Science Foundation, grant #1713752, and Harvard Business School. We thank them for their generous support of this work. We thank Ned Augenblick, Lucas Coffman, P.J. Healy, Alex Imas, Muriel Niederle, Matthew Rabin, Andrei Shleifer, Joshua Schwartzstein, Charles Sprenger, and Lise Vesterlund for their helpful feedback on this work.

[ǂ] Corresponding author: 445 Baker Library, Harvard Business School, Boston, MA; kcoffman@hbs.edu.

I. INTRODUCTION

Beliefs about own ability are key inputs into many economically significant decisions. They shape financial decision-making and educational choices, such as what schools to apply to and what fields to study. They also impact labor market outcomes, shaping how job candidates present themselves and what opportunities they apply to. In these settings, uncertainty about own ability creates space for biases, including gender biases, to flourish. In fact, across financial, educational, and professional contexts, important gender differences in beliefs have been documented (Barber and Odean 2001 on finance; Pan 2018 and Buser et al 2014 on education; Reuben et al 2014 and Exley and Kessler 2021 on self-promotion; and Coffman et al 2021 on job applications).

Across many studies, researchers have documented a gender gap in beliefs about own ability, primarily in male-typed fields, where perceived and/or actual gender gaps in performance favor men. That is, conditional on having the same measured ability, women have been found to have more pessimistic beliefs about own ability compared to men in male-typed fields. For instance, given the same ability in a number-adding task, women believe they rank worse relative to others than men do (Niederle and Vesterlund 2007). This gender gap has been found in studies that focus on "estimation", where participants estimate their own absolute performance on a task (Lundeberg, Fox, and Punćcohaŕ 1994, Deaux and Farris 1977, Pulford and Colman 1997, Beyer 1990, Beyer and Bowden 1997, Beyer 1998, Coffman 2014, Bordalo et al 2019), and in studies that ask about believed ability relative to others (like Niederle and Vesterlund 2007, and also Grosse and Reiner 2010, Dreber, Essen, and Ranehill 2011, Shurchkov 2012). These gaps in self-assessments are typically reduced or reversed in female-typed domains (Coffman 2014, Bordalo et al 2019).

Given the existence of these gender gaps and the importance of beliefs in driving decision-making, a natural question is why these gaps persist, and what might we do about them. Perhaps the most obvious suggestion for closing gaps is providing increased information.[1] If a student is unsure of her abilities in STEM, her school could provide her with more feedback about her talents in this area (based on test scores or teacher recommendations). Or, if an entry-level employee is unsure whether she possesses the qualifications needed to apply for an internal promotion, her manager could provide a more detailed performance review. If uncertainty is a driver of biases in beliefs of own ability, increased (objective) information about own ability would seem to offer a promising path toward reducing gender gaps in beliefs.

[1] Of course, another natural question is, taken these gender gaps as a given, how can we modify our processes and institutions in such a way that biased beliefs are less distortionary for outcomes? While not the focus of this paper, this is another important issue to wrestle with and is addressed in concurrent work by the authors (Coffman et al. 2021).

Our goal in this paper is to provide an empirical investigation of the effectiveness of increased information in reducing gender gaps in beliefs of own objective ability. We ask whether and how biased beliefs persist in the face of highly informative, task-specific feedback. In particular, we run a controlled experiment that explores how individuals update their beliefs about their own objective performance in response to informative, but imperfect, feedback. A central focus of our work is understanding the role of gender stereotypes in driving gender differences in reactions to this information. We know that stereotypes have important predictive power for individuals' beliefs about their own ability - absent feedback (Coffman 2014, Bordalo et al 2019). Is it also the case that these gender stereotypes have predictive power for how individuals incorporate new information into their beliefs?

Our experiment is tailored to explore whether there are gender differences in belief updating, and how these differences depend upon the gender congruence of the domain. By gender congruence, we mean the extent to which a domain carries a stereotype that favors the individual's gender: male-typed domains are more gender congruent for men, while female-typed domains are more gender congruent for women. To get at this question, we explore belief updating across eight different domains, chosen to vary in their associated gender stereotype. This allows us to ask whether reactions to information depend upon whether that information arrives in a gender congruent domain.

Participants in our experiment complete three rounds, each round featuring a different randomly-assigned domain. Within each round, participants complete a timed 20-question multiple-choice test in that domain. After completing the test, participants provide an incentivized prior belief of both their absolute performance on that test (number of questions answered correctly) and their belief of their rank relative to others completing the same test. Furthermore, we elicit a full belief distribution over all possible scores (what is the chance you answered 0 questions correctly, 1 question correctly, …, 20 questions correctly). This allow us to explore gender differences in the shape of belief distributions and to create a Bayesian benchmark for updating behavior for each participant.

Then, we provide information: a noisy signal of their true score. Across two randomly-assigned conditions, we vary the precision of the signal received. In both treatments, signals are equal to a participant's true score with probability *p* (either 0.5 or 0.7 depending on assigned treatment), and with probability *1-p,* the signal is constructed by adding an integer drawn from a uniform distribution over {-5,-4,-3,-2,-1,1,2,3,4,5} to their true score. Individuals have complete information about this signal structure. Importantly, with this signal structure, there is no selection into the quality or accuracy of the signal received: within a treatment, every individual has the same chance of receiving "good news" relative to their true performance or "bad news" relative to their true performance. This stylized feedback is highly informative, immediate, task-specific, and individualized; thus, in many ways, we have attempted to create the "best-case scenario" for the effectiveness of feedback. At the same time, our feedback in both treatments is imperfect, mimicking the uncertainty about information quality that often accompanies performance signals in the field. By comparing across two signal treatments, we can test whether more informative signals are more effective at reducing gender gaps. Finally, we collect posterior beliefs, both of absolute and relative ability, using the same belief elicitations we used for prior beliefs.

In line with past work, we find a significant role for stereotypes in predicting prior beliefs, both of absolute and relative performance. Holding fixed measured performance, individuals' beliefs of their own performance increase significantly as the category becomes more gender congruent.

Our main finding is that, after the provision of information, stereotypes continue to play a significant role in predicting beliefs. We show that the impact of stereotypes on posterior beliefs represents a systematic deviation from the Bayesian benchmark. Our rich data allows us to explore the nature of these departures from the Bayesian model. Similar to past work in related paradigms, we find that individuals update their beliefs more conservatively than the Bayesian model predicts. Our advance is showing that the extent of conservatism depends significantly on gender stereotypes. Men are significantly more responsive to information in male-typed domains, while women are significantly more responsive in female-typed domains.

This is driven by differences in reactions to positive feedback. We document how updating varies depending upon whether the signal drawn was exogenously "good" news (a signal greater than or equal to their true score) or "bad" news (a signal less than their true score). We find that individuals update more in response to good news in a gender congruent domain than in a gender incongruent domain. For example, women's beliefs increase more after seeing good news in a female-typed domain than in a male-typed domain (even relative to the Bayesian prediction). Our results suggest that convincing people of their talent in gender incongruent domains may be more challenging, as individuals seem to discount positive information in these areas more than in gender congruent domains.

Our paper contributes to a growing literature on belief updating (see Benjamin 2019 for an overview). Work on responses to performance feedback dates back to the early work of Heider (1958), who observed that individuals engage in self-serving biases when given negative feedback but are more likely to attribute positive feedback to own ability. Similarly, a meta-analysis by Campbell and Sedikides (1999) highlighted that individuals are more motivated to find outside factors to explain bad news when they are more invested in the task, for example due to high self-esteem.

Most prior studies in the economics literature have focused on paradigms that allow for the careful measurement of incentivized beliefs and clean, practical tests of Bayesian models. This often involves focusing on beliefs of relative ability. For instance, a participant might be asked their belief about the probability of placing in the top half of performers, and then receive a noisy binary signal of whether they are indeed in the top half. With this structure, one can elicit simple prior belief distributions and compare updating behavior to a full Bayesian benchmark (see, for instance, Mobius et al 2022, Barron 2016, Buser et al 2018, Coutts 2018, Gotthard-Real 2017, Ertac 2011). Eil and Rao (2011) operate in a finer response space, eliciting full belief distributions over relative placement in a population in terms of IQ and beauty (and in a non-ego-relevant control task). Within these paradigms, there is evidence that people are more confident about their probability of being among the top performers when they are motivated to be so, either because of strategic considerations (Schwardmann and van der Weele 2018) or when the task is more ego-relevant (Drobner and Goerg 2022, Buser et al 2018, Ertac 2011), consistent with theoretical models of motivated reasoning (such as Rabin and Schrag 1999, Benabou and Tirole 2002, and Koszegi 2006).

This literature has focused on conservatism – are people less responsive to information than the Bayesian model would predict? – and asymmetry – do people respond differently in response to good versus bad news? Some studies have found that individuals are more conservative, relative to the Bayesian model, when updating beliefs about themselves compared to non-self-relevant beliefs (Mobius et al 2022, Eil and Rao 2011). There is mixed evidence on asymmetry. Eil and Rao (2011), Mobius et al (2022), Charness and Dave (2017), and Zimmermann (2020) find that participants respond more to positive signals than negative signals. But, Ertac (2011) and Coutts (2017) find that participants respond more to bad news than good in their ego-relevant tasks. Other studies find no asymmetry in either direction (Grossman and Owens 2012, Buser et al 2018, Schwardmann and Van der Weele 2016, Barron 2016, and Gotthard-Real 2017).

Results on gender within this literature have also been mixed, with some studies finding gender differences in conservatism and other studies finding gender differences in asymmetry. Mobius et al (2022) report that women demonstrate more conservatism than men when updating their beliefs about the probability of placing among the top half of performers on an IQ test, updating less in response to information. But, they find no gender differences in asymmetry. Coutts (2018) documents very similar results on gender, reporting no gender differences in asymmetry and evidence of more female conservatism (in both ego-relevant and non-ego-relevant settings). On the other hand, Ertac (2011) does find some evidence of gender differences in asymmetry, with women responding less to "good news" than men do in a verbal task, but not in an addition task. Shastry, Shurchkov, and Xia (2018) also find evidence of asymmetry, showing that negative feedback disproportionately deters tournament entry for high ability women, primarily because they are too likely to attribute this feedback to ability rather than chance. Because these studies vary in their paradigms and tasks used, it is hard to know exactly what underlies any across-study differences. Building on these previous studies, one key goal of our work is to construct an environment that allows for a systematic investigation of gender differences, taking care to causally identify the roles of signal informativeness, signal valence, and domain-type in contributing to any gender differences. Perhaps most critically, we explore gender differences across a range of domains that vary in their associated gender type, allowing us to separate the impact of gender from the impact of stereotypes. This distinction is important not only for better understanding mechanisms, but also for understanding how to extrapolate to field contexts of interest.

In line with previous work on gender differences in belief updating, our project focuses on how men and women update their beliefs about objectively-measured performance in response to noisy information. These paradigms are useful for understanding how individuals form and adjust their privately held beliefs about their own talents, particularly in contexts with uncertainty. Of course, one could alternatively ask how men and women describe past performance in more subjective terms.[2] In early work, Deaux and Farris (1977) consider how men and women describe their own performances on an anagram task that the experimenters have labeled either male-typed or female-typed. When the task is described as male-typed, men evaluate their own (known) objective performance more favorably – in subjective terms - after the fact. Exley and Kessler (2022) find a very similar result, showing that even when individuals are perfectly informed about their objective performance, men are more likely to self-promote, choosing to assess their own performance more favorably. And, this gender gap is eliminated when they switch from a male-typed to a female-typed task. Our focus is not on self-promotion or subjective descriptions of performance. Instead, we investigate the ways in which stereotypes shape how individuals update their beliefs about their objective performance in response to noisy information. This enables us to compare beliefs not only across gender and stereotype, but also to an accurate beliefs benchmark and a Bayesian benchmark.

Overall, our main contribution is to document and unpack the role of stereotypes in shaping how individuals update their beliefs about themselves. We elicit incentivized prior and posterior beliefs of performance, in an environment with uncertainty about true performance. Within a single controlled study, we compare men and women across a range of domains, and ask how, holding other factors fixed, the gender-type of the environment matters for decision-making. To do so, we systematically vary the gender-type of the domain comparing a range of female-typed tasks to a range of male-typed tasks, taking care to balance difficulty and format to better isolate the stereotype component.

[2] We refer to an assessment as subjective if there is no objectively-defined "right" answer. Consider an example. Suppose we asked someone to guess their GPA from last semester. This is a belief about an objective measure of performance; we could incentivize a truthful guess because there is a clear, objective benchmark. If we instead asked someone, given your (known) GPA, how would you assess your academic performance last semester (for instance, on 1 – 7 scale where 7 is excellent), this would be a subjective assessment; we could not incentivize a truthful guess because there is no clear, objective benchmark for how GPA *should* translate into the subjective assessment. Our paper focuses on beliefs about objective measures of performance.

In addition, by focusing on absolute ability rather than relative ability, our experiment will yield rich, well-identified data on asymmetry. Due to our signal structure, men and women across the ability spectrum will be equally likely to receive good or bad news (of equal accuracy) relative to true performance. Because participants update on absolute ability, the space of possible beliefs will be quite fine, potentially allowing for identification of more subtle differences.[3] Our results suggest that past findings of greater female conservatism could possibly be explained by (i) a sampling of primarily more male-typed domains, and/or (ii) an under-appreciation of gender differences in variance in priors. Our framework also helps provide additional insights into the mixed results on gender differences in asymmetry. Our results suggest that responsiveness to good news, in particular, is a function of how gender congruent the domain is, not simply a function of gender.

Across both educational and professional contexts, individuals regularly receive feedback on their own abilities. This information, even if unbiased, will almost always be noisy relative to the true object of interest. In this way, our experimental framework asks a question that is central to understanding the evolution of beliefs over time: how does new, highly informative (but imperfect) feedback shape beliefs of own ability? Our results suggest that policy interventions aimed at closing gender gaps in self-confidence that simply provide feedback to individuals may not have as strong of an impact as intuition or the Bayesian model would predict. Rather, gender stereotypes seem to impact the way new information is incorporated into beliefs, fueling persistence in gender gaps.

II. EXPERIMENTAL DESIGN

Our controlled experiment features three rounds for each participant. Within each round, participants complete a test, provide prior beliefs of their performance, receive noisy feedback on their performance, and provide posterior beliefs. Across each round, we randomly vary the domain of the test. Building on the design of Coffman (2014) and Bordalo et al (2019), we select eight different domains that vary in their associated gender stereotype: Cars, Sports, Videogames, Business, Verbal Skills, Art and Literature, Disney Movies, and Kardashians.[4] While some of these categories lack career or educational relevance, their clear associated gender-types allow for well-powered identification of the role of stereotypes in driving beliefs. In Appendix A, we show that the patterns we obtain for the male-typed domains in this study are similar to the patterns we observe in a male-typed domain with clear external relevance: cognitive skills. We also

[3] This is closest in design to the work of Eil and Rao (2011), who focus on relative ability but allow for belief distributions over all possible ranks in a population. Gender is not a focus of their study.

[4] The questions and domains are drawn from those used by Bordalo et al (2019). We document the associated gender stereotypes in Figure I in the hypotheses and empirical approach section.

show that our results are robust to considering only the more academically relevant domains of Business, Verbal Skills, and Art and Literature. Below, we describe each stage of the round in detail.

Timed Ability Test

For each of the eight domains, we construct a 20-question multiple-choice test. Each multiple-choice test is a timed test, where participants are awarded 1 point for each correct answer. Skipped or incorrect answers are not penalized. Participants have three minutes to answer as many questions as they can, and receive $0.25 for each correct answer if that round of the experiment is selected for payment.

Prior Beliefs

Following their completion of the test, we elicit beliefs from participants. Each belief question is incentivized. First, we ask the participant what they think their most likely score on the test was. This is incentivized by offering a bonus payment if they guess their score exactly correctly. The advantage of this elicitation is that it is simple and intuitive for participants, likely capturing meaningful information about their beliefs. However, if we want to understand how a Bayesian would update their beliefs in response to the feedback we provide, this belief measure is insufficient; we need to understand their full subjective belief distribution over possible scores. We try to collect this complex information in a way that is still intuitive for participants.

After reporting what they believe to be their most likely score, on the next page participants are asked their perceived likelihood that they earned this exact score. For example, suppose a participant guessed they had a score of "6," they would be asked to complete the sentence: "I believe there is a __% chance I earned exactly a score of 6." This question elicits the probability mass they assign to the mode of their belief distribution. Then, on the next page, we then ask them for their full subjective distribution over all possible scores, reminding them of the probability mass that they assigned to the mode of their prior. We impose that the probabilities sum to one.

Finally, we measure believed relative ability. We ask participants what their believed rank is, comparing themselves to 100 other randomly-chosen participants who completed the same multiple-choice test. They can guess any particular rank between 1 – 100. We incentivize them to report the mode of their prior over all possible ranks by paying them a bonus payment if they guess their exact rank correctly.

Provision of Signals

After providing their prior beliefs, participants receive a noisy signal of their performance on the test. With probability *p,* the signal transmitted is exactly equal to their score on the test. With probability *1 – p*, the signal is equal to their score plus randomly-drawn noise. The noise is drawn from a uniform distribution over non-zero integers between -5 and 5, that is: {-5, -4, -3, -2, -1, 1, 2, 3, 4, 5}.

At the outset of the experiment, participants are randomly assigned to one of two signal treatments, either the *50% Signal Treatment* (*p* = 0.5) or the *70% Signal Treatment* (*p* = 0.7). This determines the noise structure of the performance signal they receive across all three rounds. By varying *p* across participant, we can explore whether more informative signals are more effective at reducing the extent to which gender stereotypes shape posterior beliefs.

We explain the signal mechanism to participants in detail. They are told to imagine 10 balls, numbered 1 – 10, in a bag. The computer will draw one of those balls at random. If the computer draws a ball with a number between 1 – 5 (or 1 - 7 for those in the 70% Signal Treatment), the computer will show them their true test score. But, if the computer draws a number between 6 – 10 (or between 8 - 10 in the 70% Signal Treatment), the computer will show them their true score plus some error, where the error is equally likely to be any non-zero integer between -5 and 5. That is, the computer will take their score and add either -5, -4, -3, -2, -1, 1, 2, 3, 4, or 5 to construct their signal.

We tell them explicitly that they will just see their signal, not what ball the computer chose, nor what error the computer chose. We then give them a few examples of how different scores, draws of balls from the bag, and errors would produce different signals. We close by emphasizing that the computer will show them their true score 50% (70%) of the time. They then answer a brief understanding question that they must answer correctly before continuing.

Elicitation of Posterior Beliefs

After they see their signal, participants provide posterior beliefs. We re-ask all the beliefs questions, including their believed most likely score, the likelihood they associate with this particular score, their full beliefs distribution over all possible scores, and their believed rank compared to 100 other participants. Participants receive no additional feedback before completing the next round of the experiment. The next round of the experiment is identical, except that they see a new, randomly-drawn domain.

Follow-up Questions

Following the three rounds of the experiment, participants complete a brief demographic questionnaire that asks their gender, race, educational attainment, and whether or not they attended high school in the United States. We also include five unincentivized cognitive skills questions, performance on which we use as a control variable when predicting beliefs. Finally, the very last question of the experiment asks them about the believed gender stereotype they associate with each of the eight possible domains in the experiment. They are given a slider scale that ranges from – 1 (women know much more) to 1 (men know much more) and are asked to indicate, using the slider scale, which gender on average they believe knows more about each domain.

Implementation

We conducted the study on Amazon Mechanical Turk with 2,025 participants (25 of which participated one day ahead of the full HIT to ensure the functionality of the programming) in October 2018. We take a number of steps to ensure data quality. The study was restricted to workers with a United States based IP address who had completed at least 100 tasks (called HITs) and had an approval rating by previous requesters of at least 95%. We include understanding questions and attention checks as well as a captcha to screen out bots. Participants must answer the understanding questions correctly in order to complete the study. Attention checks were presented to a random sub-sample of participants; the attention check requires the participant to select the correct picture within seven seconds (i.e., identify the picture of blueberries on this screen). A participant fails the attention check if they select the wrong picture or if they did not select any picture within the seven seconds given to them. Of the 770 participants who viewed the first attention check, 98% pass it; of the 771 participants who saw the second, 97% pass it. We exclude the 36 participants who failed either attention check, leaving us with 1,989 participants.

The HIT was advertised as a 30-minute academic study that guaranteed a completion payment of $2.00 plus the possibility of incentive pay. Participants were told that one domain would be chosen at random to determine their bonus payment. For this randomly-selected round, they received $0.25 per problem solved correctly on the multiple-choice test. In addition, for all beliefs questions asked within the round, one was chosen at random as the “decision-that-counts”. If the decision that counted was their believed score or believed rank, they received $0.50 if they guessed correctly. If the decision that counted was instead about the probability mass they assigned to a particular score, we used an adaptation of a BDM to incentivize truthful reporting, independent of risk preferences. All participants were told that we were incentivizing them to tell the truth. They also had the option of clicking on a link that said “Here is why you should tell the truth” that explained the procedure in detail.[5]

[5] We use a variation of the procedure in Mobius et al (2022) for eliciting subjective beliefs. In particular, suppose we are eliciting the participant’s believed likelihood that they earned a score of “7” on a particular test. Participants are told that there are 100 different “random number” lotteries; each “random number” lottery pays the prize with X% chance and pays nothing with (1-X)% chance. The 100 lotteries vary X, increasing it from 1, to 2, to 3, all the way to 100. By telling us their perceived probability of earning exactly a “7” on the test, participants are communicating to us which “random number” lotteries they prefer to a simple bet on their score. The bet on their score pays the prize if

Note that during the running of the experiment, we noticed that there was an error in the specific click-through instructions available to participants describing how truth-telling was incentivized for probability distribution questions. While the language still emphasized truth-telling as the expected money-maximizing strategy, the specific incentives were described incorrectly. This error was corrected in the middle of the experiment, and a comparison of participant answers before and after the error correction does not suggest that the error impacted the responses given. A full analysis of this issue is presented in Appendix C. The complete experimental materials are provided in Online Appendix F.

III. HYPOTHESES AND EMPIRICAL APPROACH

Our goal is to understand what gender gaps in beliefs look like, conditional on performance, and what role stereotypes play in predicting these gender differences. To formally measure the impact of stereotypes, we follow the approach of Bordalo et al (2019). Under this model, a decision-maker's belief about herself is shaped, in part, by comparisons of the performance of her own gender in a category compared to the performance of the opposite gender. Beliefs about own performance are then exaggerated in the direction of true gender gaps. That is, holding own individual ability fixed, the model hypothesizes that women's (men's) beliefs about own performance will increase as the average female (male) advantage in a category increases. In this way, stereotypes produce gender gaps in beliefs that are larger than (but directionally in line with) true gender gaps in performance.

We designed the experiment to produce the variation in gender stereotypes across domain that is needed to estimate this model. Appendix Table B1 documents that our categories vary significantly both in terms of observed gender gaps in performance and perceived gender gaps in ability. Figure I illustrates the key data. We arrange the categories by the average slider scale rating given for the category among all participants, a measure of perceived male advantage. This average slider scale rating is graphed as the black line against the secondary y-axis. Four categories are perceived as being female-typed: Kardashians, Disney, Art and Literature, and Verbal Skills, while four categories are perceived as being male-typed: Business, Videogames, Sports, and Cars. We will use these classifications when we refer to female-typed or male-typed domains going forward. The bar graph illustrates the average male and female score in each domain.

they indeed have the score of "7," and pays nothing otherwise. We draw a random number lottery at random from the set of 100. If the chance of that random number lottery paying the prize (X) is greater than the participant's stated belief, then they are paid based upon the outcome of that random number lottery drawing. If, on the other hand, the chance of that random number lottery paying the prize (X) is less than the participant's stated belief, then they are paid based upon the bet on their score – earning the prize if indeed their score on that test was "7." The full language used to explain the procedure is available in Appendix F.

The average gender gaps in performance correspond quite closely to the slider scale perceptions. In fact, if we correlate the male advantage in performance within a domain (gender gap in average test scores) with the average slider scale rating of that domain provided by participants, the correlation is 0.88. Note that there is heterogeneity in domain difficulty (average performance) within and across gender-type. For instance, women's performance in videogames exceeds their performance in verbal skills, and men's performance in art exceeds their performance in business. This is important as we want to separate the impact of the gender stereotype – perceptions of relative differences across gender – from the direct impact of category difficulty.

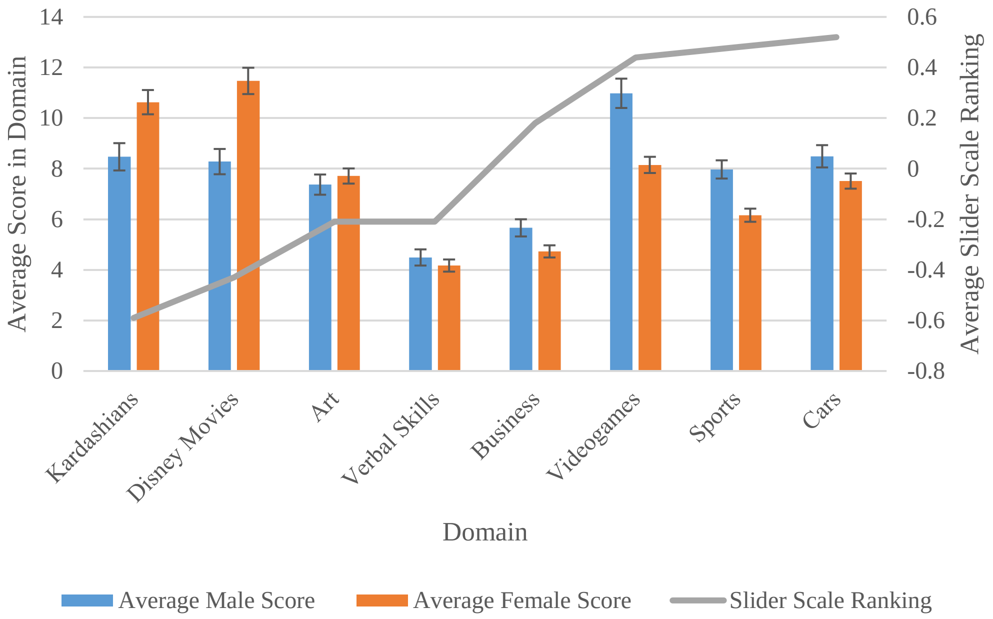


Figure I. Variation in Gender-Type of Domains

Error bars illustrate 95% confidence intervals.

Following Bordalo et al (2019), we can test for stereotyping by exploring whether the gender gap in average performance within a category is predictive of an individual's belief about herself, holding fixed her own measured performance. The hypothesis is that as own gender advantage within a category increases, an individual will report more optimistic beliefs about her own performance, even holding fixed her own performance. This suggests the following specification:

$$(1)\ Belief_{i,j,t} = B_0 + B_1(Female) + B_2\left(\overline{Score}_{j,G} - \overline{Score}_{j,-G}\right) + B_3(Score_{i,j})$$

where $Belief_{i,j,t}$ represents the reported belief of individual i in category j at time t (prior or posterior), *Female* is an indicator variable capturing the participant's gender, $\left(\overline{Score}_{j,G} - \overline{Score}_{j,-G}\right)$ is the difference in average scores in category j between members of the participant's own gender, G, and the opposite gender, -G, and $Score_{i,j}$ represents the score of individual *i* in category *j*.

Gillen, Snowberg, and Yariv (2019) point out that measurement error in control variables can bias coefficients of interest. Here, the main concern is that test score is only a noisy measure of ability, and unobserved ability may be correlated with gender or own gender advantage. So, following the recommendation of Gillen et al (2019), we can add to our specifications a second proxy for individual performance: the average performance of one's own gender in the category.

$$(2)\ Belief_{i,j,t} = B_0 + B_1(Female_i) + B_2\left(\overline{Score}_{j,G} - \overline{Score}_{j,-G}\right) + B_3\left(Score_{i,j}\right) + B_4\left(\overline{Score}_{j,G}\right)$$

By including average performance of one's own gender in our regressions, we reduce concerns about measurement error and limit the chances that the gender *gap* is informative for beliefs simply because it is proxying for gender-specific performance in that domain more broadly.

Before continuing, we offer a few comments on the inputs to this model. In our main analysis, we use a participant's guess of her score as our measure of $Belief_{i,j,t}$. This measure is easy to understand, incentivized simply, and is straightforwardly related to our signal of absolute performance. For ease, we use "belief" to refer to this guess, acknowledging that we have other beliefs measures we could analyze. In Appendix B, we document the robustness of our results to using another measure of beliefs about absolute performance, the mean of the full belief distribution provided by the participant.

Importantly, we control for a participant's score in the domain. Similar to past work, we observe that the degree of underestimation increases with performance, both for men and women (see Lichtenstein and Fischhoff 1977, Moore and Healy 2008, and Bordalo et al 2019).This trend makes it essential that we account for observed performance when comparing the degree of confidence across men and women, so that we can isolate the role of gender from the role of difficulty of the task in shaping beliefs. A simple comparison of means (of performance and of beliefs) confounds these two distinct factors.[6] We use a linear control for score in our main analysis and present robustness checks in Appendix B. Our results are qualitatively robust to more flexible approaches, such as fixed effects for each score, though estimated effect sizes are reduced by approximately 1/3.

[6] For instance, we expect that the role of gender is such that women will under-estimate performance relative to men (given the same observed performance). But, we also expect (and observe) that lower-performing participants will over-estimate scores more compared to higher-performing participants. Because women are lower-performing than men on average in male-typed domains, these two effects can offset or mask one another. Therefore, it is important to control for performance.

We control for the participant's gender, allowing us to identify any gap in beliefs between men and women on average across the domains. We also include our other demographic controls (whether or not she attended high school in the U.S., fixed effects for educational attainment, fixed effects for race, her score on the five cognitive ability questions), and round fixed effects. Omitting demographic controls from the specification does not alter our results. We cluster standard errors at the individual level.

The gender gap in average performance in the category, signed to match the participant's gender, is our proxy for the associated stereotype of the domain. The coefficient on this difference is our key coefficient of interest, B_2, the impact of own gender advantage on beliefs. This model provides the foundation for exploring the impact of stereotypes on beliefs. We can estimate this model first on prior beliefs and then on posterior beliefs. By comparing B_2 across the two models, we can ask whether the impact of stereotyping on beliefs persists after the provision of feedback.

Question 1: Do gender stereotypes predict beliefs after the provision of noisy feedback?

We can also leverage the difference in signal accuracy between our two treatments to ask whether more informative signals are more effective at reducing the impact of gender stereotypes. We test Question 1 separately for individuals randomly assigned to the *50% Signal Accuracy* treatment and for those assigned to the 7*0% Signal Accuracy* treatment. This comparison speaks to the optimal design of feedback outside of this study; if noisier feedback is more likely to perpetuate gender gaps, this suggests a need to increase the informativeness of feedback interventions. If, on the other hand, more informative feedback is no more effective in closing gender gaps, this suggests a need to consider other types of interventions.

Question 1a: Does the impact of stereotyping on posterior beliefs depend upon the informativeness of the noisy feedback?

The Bayesian model has been a foundation of previous work on belief updating, with past papers exploring whether men and women vary in the extent to which they display conservativeness or asymmetry relative to this classic model. Our wealth of data allows us to ask the critical question of whether gender stereotypes predict deviations from the Bayesian model.

For each participant-domain observation in our dataset, we can construct a Bayesian-predicted posterior, using the participant's subjective prior and the signal she received. The key insight is that, under our signal structure, the Bayesian prediction for most participants is of one of two types: her guess of her score should be the signal she observed, or her guess should be the same as her prior guess of her performance. Oversimplifying – either believe the signal or ignore it. The intuition is as follows: signals are accurate enough in our setting that, if a participant assigned sufficient probability to the signal in her prior belief distribution, she should update her guess of her score to that signal after observing it. If she did not put sufficient weight on the signal ex ante, then she should continue to report her prior guess as her posterior guess. In Appendix D, we formalize this intuition, developing propositions that characterize the Bayesian prediction for each participant.

For a small fraction of our participants, the Bayesian prediction is unclear: these are the participants who put no positive prior probability on any score that could generate the signal in their prior belief distribution.[7] These participants have seen a zero probability event. This occurs for 8% of men's observations and 10% of women's observations. We make the assumption in the analysis below that the most reasonable Bayesian prediction for these participants is to report their signal – this is the belief that would be justified by any (non-zero) flat prior over the scores that could generate the signal.[8]

Overall, for 54% of observations from men and 50% of observations from women, the Bayesian prediction is that the participant should report her signal as her posterior belief. For 47% of observations from men and 50% of observations from women, the Bayesian prediction is that the participant should report her prior belief as her posterior belief. Finally, for 10% of observations from men and women, the Bayesian prediction is some other score (not the signal nor the prior guess) – this occurs only in the cases where the mode of the prior could not have generated the signal observed *and* there is sufficiently little weight on the signal in the prior distribution, *and* there is some non-zero weight on a score that could have generated the signal (see Proposition 2 in Appendix D).[9] Note that these proportions do not sum to 100% because for some participants, the signal is their prior guess.

Returning to our empirical approach, we can insert these Bayesian predictions into the model we proposed earlier, asking whether stereotypes predict deviations from the Bayesian benchmark:

[7] Note that the large majority of our participants receive a plausible signal. For 80% of observations, a true score equal to the mode of the believed distribution could generate the observed signal. For 91% of observations, the participant put positive prior probability on at least one score that could have generated the signal. And, only 159 signals fall outside of the 0 – 20 range. For completeness, we include all observations in our main text analysis.

[8] More formally, we can account for this issue by smoothing participant priors. That is, for each person-domain observation in the dataset, we modify their prior by adding a small positive probability (1 percent) to each score that received zero probability in the original prior. We then re-scale the prior proportionally so that probabilities sum to 100 percent. With these smoothed priors, there is a well-defined Bayesian prediction for all observations. In the Appendix, we show that our results are unchanged if we follow this alternative approach. In fact, for 95% of observations, the Bayesian prediction is the same under these two procedures.

[9] For these participants, the Bayesian prediction is that the participant report the mode of her prior restricted to the set of scores that could have generated the signal. For those participants who have multiple modes in this space (173 of the 572 observations in this group), we take the average of those modes as the Bayesian prediction.

$$(3)\ Belief_{i,j,t} = B_0 + B_1(Female_i) + B_2\left(\overline{Score}_{j,G} - \overline{Score}_{j,-G}\right) + B_3\left(Score_{i,j}\right) + B_4\left(\overline{Score}_{j,G}\right) + B_5(Bayes_{i,j})$$

where $Bayes_{i,j}$ is simply the Bayesian predicted posterior belief for individual *i* in category *j*. This allows for a test of our next question.

Question 2: Do gender stereotypes predict posterior beliefs, even conditional on the Bayesian-predicted posterior?

Suppose we cannot reject that $B_2 = 0$ in this updated model; that is, conditional on the Bayesian-predicted posterior, gender stereotypes have no significant impact on posterior beliefs. This would reveal that any impact of stereotypes on posterior beliefs is well-explained by the Bayesian model. If instead we estimate $B_2 > 0$, this would tell us that stereotypes impact posterior beliefs over and above how the Bayesian model would predict. We would conclude that stereotypes produce non-Bayesian updating of beliefs in response to new information.

Past work on belief updating has focused on two types of non-Bayesian behavior: conservativism and asymmetry. We explore how the non-Bayesian impact of stereotypes on posterior beliefs relates to these two commonly studied deviations, connecting with past work and better understanding gender differences in these behaviors. While our first empirical approach was geared toward isolating the role of gender stereotypes in predicting beliefs, accounting for any potential confounds, our next empirical strategy serves a different goal: produce easily interpretable estimates of the extent of conservatism and asymmetry that we can compare across gender and domain. To do so, we follow the approach of Eil and Rao (2011), predicting posterior beliefs from only the Bayesian-predicted posterior:

$$(4)\ Belief_{i,j,t} = B_0 + B_5(Bayes_{i,j})$$

In this simple model, B_5 provides a measure of the responsiveness of posterior beliefs; $B_5 < 1$ reveals conservatism in the extent to which individuals are updating their beliefs in response to new information. We can estimate this model separately for men and women and separately for male and female-typed domains, asking whether the extent of conservatism varies by gender or gender-type of the domain.

Question 3: Are there gender differences in conservatism? How do these differences depend upon the gender-type of the domain?

Conservatism measures the extent to which individuals under-react to information. A reasonable hypothesis is that these patterns vary depending upon the type of information received. We hypothesize that individuals may respond more to “stereotype-consistent” news. That is, we expect individuals to react more to good news in gender congruent domains, relative to incongruent domains. At the same time, we expect individuals to react *less* to bad news in gender congruent domains, relative to incongruent domains. Thus, whether the news is good or bad should matter for determining whether individuals are more or less conservative in gender congruent domains.

We can build on model (4) to explore this asymmetry: the extent of responsiveness to "good news" versus "bad news" depending on the gender-type of the domain. In our analysis, we refer to a signal as good news if it is equal to or above their true score, and we refer to a signal as bad news if it is below their true score.[10] Because the signal displayed is exogenous conditional on performance, this definition of news avoids any selection on priors or performance. Under this definition, it is **not** more likely that an under-confident participant receives good news than an overconfident participants receives good news; nor is it more likely that a talented participant receives good news than a poor-performing participant receives good news. Of course, these definitions of good and bad news may be a step removed from the participant's actual perception of whether the news is good or bad, which seems more likely to be defined relative to their prior beliefs. In this way, our analysis is like an intent-to-treat.[11] We add to our basic model an indicator for good news (signal greater than or equal to score) and the interaction of good news with the Bayesian prediction:

$$(5)\ Belief_{i,j,t} = B_0 + B_5\left(Bayes_{i,j}\right) + B_6\left(GoodNews_{i,j}\right) + B_7\left(Bayes_{i,j}\ x\ GoodNews_{i,j}\right)$$

In this specification, $B_6 > 0$ would reveal what Eil and Rao (2011) refer to as "generalized optimism" – beliefs that are on average greater than what the Bayesian model would predict after good news, relative to when the same Bayesian prediction is made for bad news. Differential responsiveness for good news relative to bad is revealed by B_7 : $B_7 > 0$ points to greater responsiveness to good news (measured by $B_5 + B_7$) relative to bad (measured by B_5). By estimating this model separately by gender and gender-type of the domain, we look for evidence of more responsiveness to "stereotype-consistent" news (good news in a congruent domain, bad news in an incongruent domain).

Question 4a: Are there gender differences in responses to good news? Do these differences depend upon the gender-type of the domain?

Question 4b: Are there gender differences in responses to bad news? Do these differences depend upon the gender-type of the domain?

[10] We choose to include truthful news as "good news" since it is more likely that a truthful draw is greater than a participant's prior guess (61% of cases) then below a participant's prior guess (25% of cases).

[11] In the Appendix, we perform the same analysis but instead define good and bad news relative to priors and find, in general, quite similar results.

Our analysis follows this empirical approach. We first document the predictive power of gender stereotypes for posterior beliefs. Then, we show that gender stereotypes have an impact on posterior beliefs beyond what the Bayesian model would predict, documenting systematic deviations. Relying on the Bayesian benchmark, we show that our participants are conservative in their reactions to information. Importantly, the extent of conservatism depends on the stereotype of the domain. Men are significantly less conservative than women in male-typed domains, while women are significantly less conservative than men in male-typed domains. Both men and women are more responsive to information in gender congruent domains. Finally, we show that these differences are driven by responses to good news. Men are significantly more responsive to good news in male-typed domains compared to women. However, this gap is reversed in female-typed domains. Both men and women are more responsive to good news when it arrives in a gender congruent domain. Taken together, our analysis suggests an important role for gender stereotypes in predicting how individuals incorporate new information into their beliefs about themselves.

IV. RESULTS

Appendix Table B1 presents summary statistics for our participants as well as an overview of the raw data. We see signs of effort and attentiveness among our participants. Appendix Figure B1 shows that the majority of participants use the full time allotted for the quiz, consistent with sincere attempts at answering as many questions as possible rather than simply clicking through. Appendix Figure B2 illustrates the full distributions of scores by gender and category.

Figure II provides a detailed picture of the beliefs data. The top panel presents the data for men, showing the distributions of both prior beliefs (solid lines) and posterior beliefs (dashed lines) conditional on each possible score. The data is split according to gender congruent (male-typed for men) categories on the green, left-hand side of the violin, and gender incongruent (female-typed for men) categories on the orange, right-hand side. To facilitate comparisons, we also present the gender-specific mean prior belief (hollow circle) and gender-specific mean posterior belief (hollow triangle) for each score. For reference, we graph the 45 degree line as a solid black line; accurate beliefs would lie on this diagonal. The second panel presents this same analysis for women.

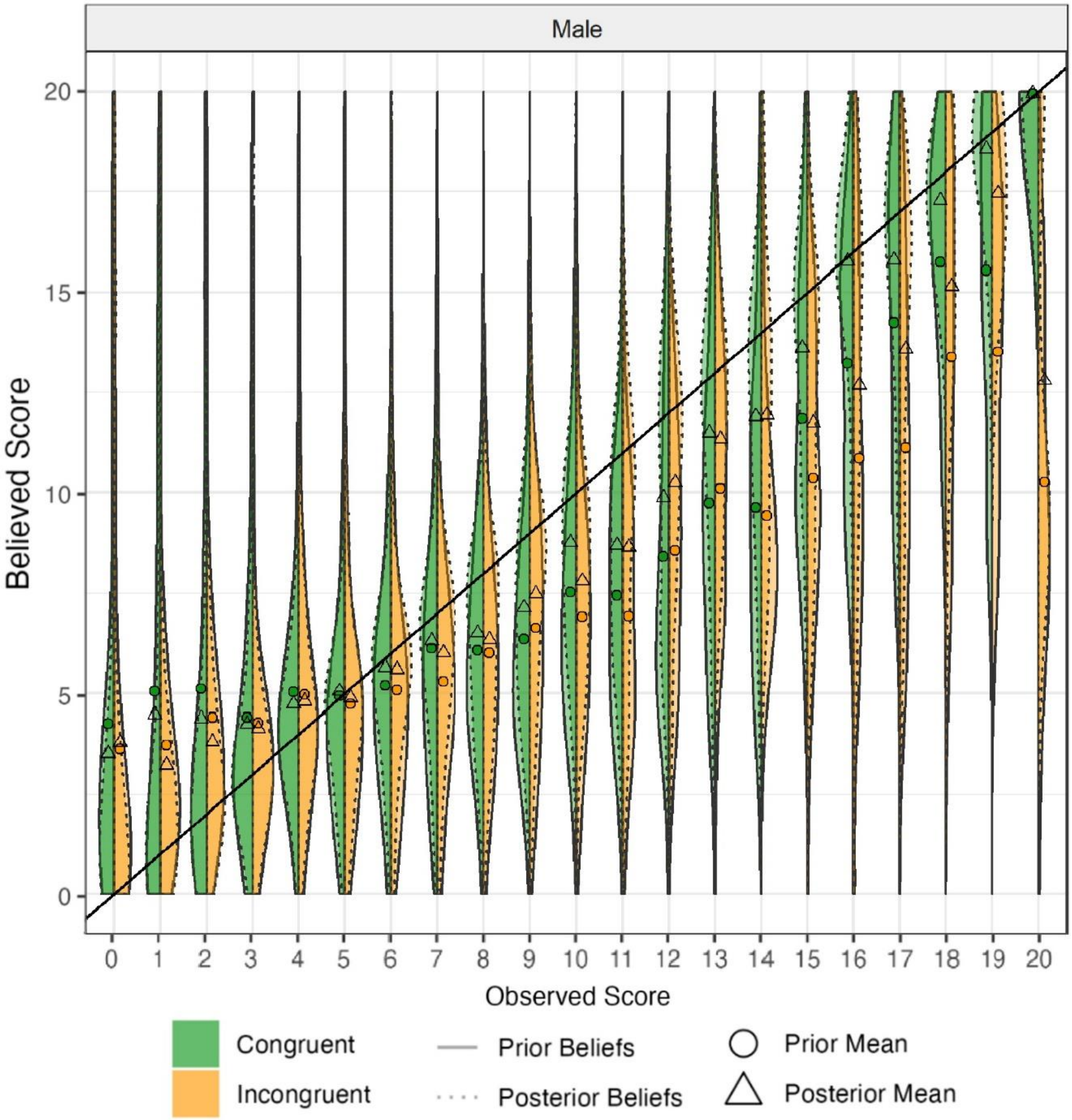


Figure IIa. Visualization of Raw Beliefs - Men

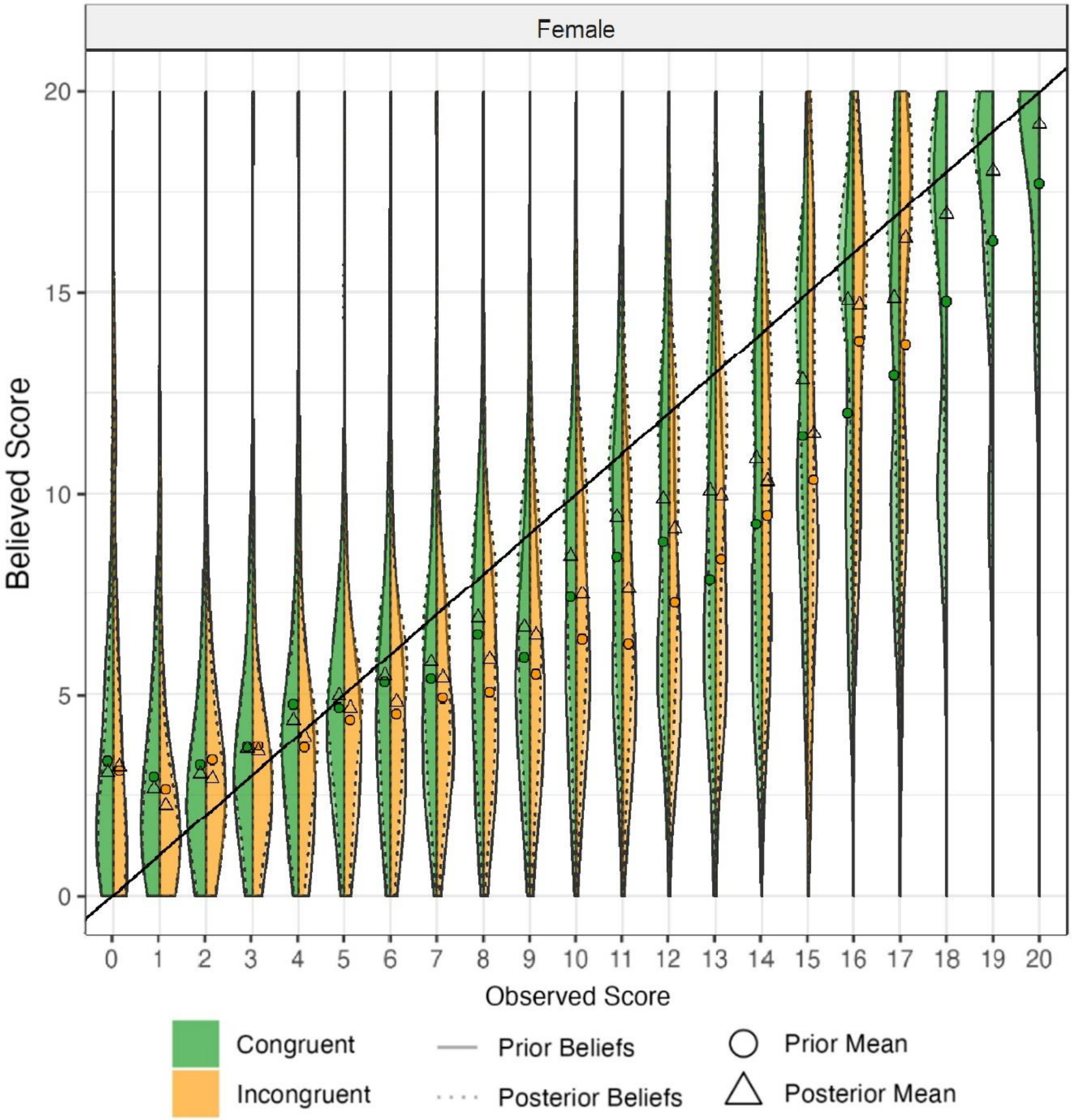


Figure IIb. Visualization of Raw Beliefs - Women

There are a number of important observations. First, beliefs of score correlate with observed score, suggesting attentiveness and understanding among our participants. Second, individuals under-estimate their scores on average, with the extent of this under-confidence increasing in observed performance. Third, when we consider prior beliefs, we notice that individuals (both men and women) are typically more confident in gender congruent than gender incongruent categories, conditional on observed performance, consistent with Coffman (2014) and Bordalo et al (2019).

But, what happens after feedback? We see that feedback seems to help move people in the direction of the truth; posteriors are less under-confident on average than priors. And, the shape of the distributions illustrate more "bunching" on accurate beliefs, consistent with individuals understanding and responding to the signals. However, and most centrally, feedback does not seem to effectively close the gaps between beliefs in congruent and incongruent categories. Even in terms of posterior beliefs, individuals appear to be significantly more confident in congruent compared to incongruent categories. This suggests a persistence of self-stereotyping in the face of informative feedback. Our formal analysis will unpack these findings.

The Impact of Gender Stereotypes on Prior and Posterior Beliefs

Table I presents the results of estimating Equation (2), first assessing prior beliefs and then assessing posterior beliefs. Column I considers prior beliefs, formally documenting self-stereotyping in beliefs prior to feedback. We estimate that a 1-point increase in male advantage in average performance decreases women's beliefs about their own ability by 0.29 points, while increasing men's beliefs about their own ability by 0.29 points.[12] If we consider the total impact of moving from a category with no gender gap to a category with a 1-point male advantage (roughly the size of the male advantage in business), we estimate an increase of 0.58 points (0.29 x 2) in the gender gap in beliefs of own ability. For reference, this is approximately 0.21 SDs of average performance in business.[13] We also estimate that, holding own and gender-specific performance fixed, women report prior beliefs approximately 0.46 points lower than men's on average, conditional on the gender-type of the category.

In Appendix Table B3, we offer analysis of the shape of the prior distributions provided by participants, with a focus on variance and skewness. We define the range of the prior as the maximum score allotted positive probability in the prior minus the minimum score allotted positive probability in the prior. Many participants provide quite tight ranges – the median range is 4 and the mean range is 5.02 (see Appendix

[12] The empirical model forces the same estimated stereotyping effect for men and women. If we instead run a model that allows for women and men to vary in their degree of self-stereotyping, we estimate a coefficient on own gender advantage of 0.21 for men (p<0.001) and 0.41 for women (significantly greater than coefficient for men, p<0.01 on the difference). We rely on the simpler model to focus on our main research questions of interest.

[13] Other approaches yield similar results (see Appendix Table B2). For instance, if instead of using the participant's guess of her score as the dependent variable we predicted the mean of her prior belief by computing the weighted average of her subjective belief distribution over all possible scores, the results are nearly identical. We also present results using a different approach to measuring the role of stereotypes, implementing the approach of Coffman (2014) and asking whether the average perception of the category as measured by the slider scale has predictive power for beliefs conditional on own measured ability. Essentially, we replace own gender advantage with own gender average perception (re-coding the sliding scales for women so that positive numbers always indicate an average perception in favor of own gender). Again, we see very similar results, with a strong estimated impact of own gender average perception of 1.18 points (p<0.001). That is, we estimate that moving from a gender-neutral category to a category that is perceived as 0.20 points on the slider scale toward male-typed (roughly the average rating of business), decreases women's beliefs of their own ability by 0.24 points and increases men's beliefs of their own ability by 0.24 points. We also show a significant effect of self-stereotyping on prior beliefs of relative performance.

Figure B3 for a histogram of the ranges). We define a "symmetric" bucket of distributions in which the mean of the distribution is also the median, a left-skewed bucket in which the median exceeds the mean, and a right-skewed bucket in which the mean exceeds the median. Approximately 21% of distributions are symmetric, 39% are left-skewed, and 40% are right-skewed. Overall, we see that women provide narrower, less variant, and slightly more right-skewed priors than men. We see little impact of the gender-type of the category on the shape of prior elicited, though participants do assign more probability mass to the mode of the prior for more gender congruent categories on average. Note that, holding all else fixed, this would make them *less* inclined to update their beliefs in response to the signal in more gender congruent categories.

Table I. Self-Stereotyping in Prior and Posterior Beliefs

	OLS Predicting Belief of Score $Belief_{i,j,t}$			
	Prior Belief	Posterior Belief		
	Full Sample	Full Sample	50% Signal Accuracy	70% Signal Accuracy
	I	II	III	IV
$Female_i$	-0.45***	-0.33***	-0.18	-0.46***
	(0.099)	(0.092)	(0.12)	(0.14)
Own Gender Advantage $\left(\overline{Score}_{j,G} - \overline{Score}_{j,-G}\right)$	0.29***	0.24***	0.24***	0.24***
	(0.024)	(0.022)	(0.031)	(0.032)
$Score_{i,j}$	0.61***	0.79***	0.79***	0.80***
	(0.013)	(0.012)	(0.016)	(0.018)
Controls	Yes	Yes	Yes	Yes
Adjusted R-squared	0.444	0.618	0.622	0.617
Clusters (Obs.)	1989 (5967)	1989 (5967)	992 (2976)	997 (2991)

Notes: * indicates significance at p<0.10, ** at p<0.05, *** at p<0.01, and **** at p<0.001. Controls are a dummy for whether participant attended high school in the United States, dummies for each race category, dummies for each education category, fixed effects for round, average score in the category for the participant's own gender, and number of ASVAB questions answered correctly.

Do signals reduce the gender gap in beliefs, and in particular the reliance on stereotypes in shaping beliefs about own ability? Table I, Column II replicates Column I but instead predicts posterior beliefs. Directionally, the reliance on stereotypes shrinks in posterior beliefs compared to prior beliefs. However, it remains sizable and significant. While a 1-point increase in own gender advantage increases beliefs of own ability by an estimated 0.29 points in priors, the same increase in own gender advantage increases beliefs of own ability in posteriors by 0.24 points.[14]

Result 1: Gender stereotypes continue to predict beliefs after the provision of noisy feedback.

Next, we ask whether more informative signals are more effective in reducing gender gaps: do we see less reliance on gender stereotypes in the 70% signal accuracy treatment than in the 50% signal accuracy treatment? Column III analyzes the 50% signal accuracy treatment and Column IV analyzes the 70% signal accuracy treatment. The results are very similar across the treatments: the estimated coefficient on own gender advantage is 0.24 in both sub-samples. More informative feedback is no more effective in reducing the impact of gender stereotypes. Given the similarity of the results across the two signal treatments, we will consider them jointly in our remaining analysis.

Result 1a: More informative signals are not significantly more effective in reducing the impact of stereotypes on posterior beliefs.

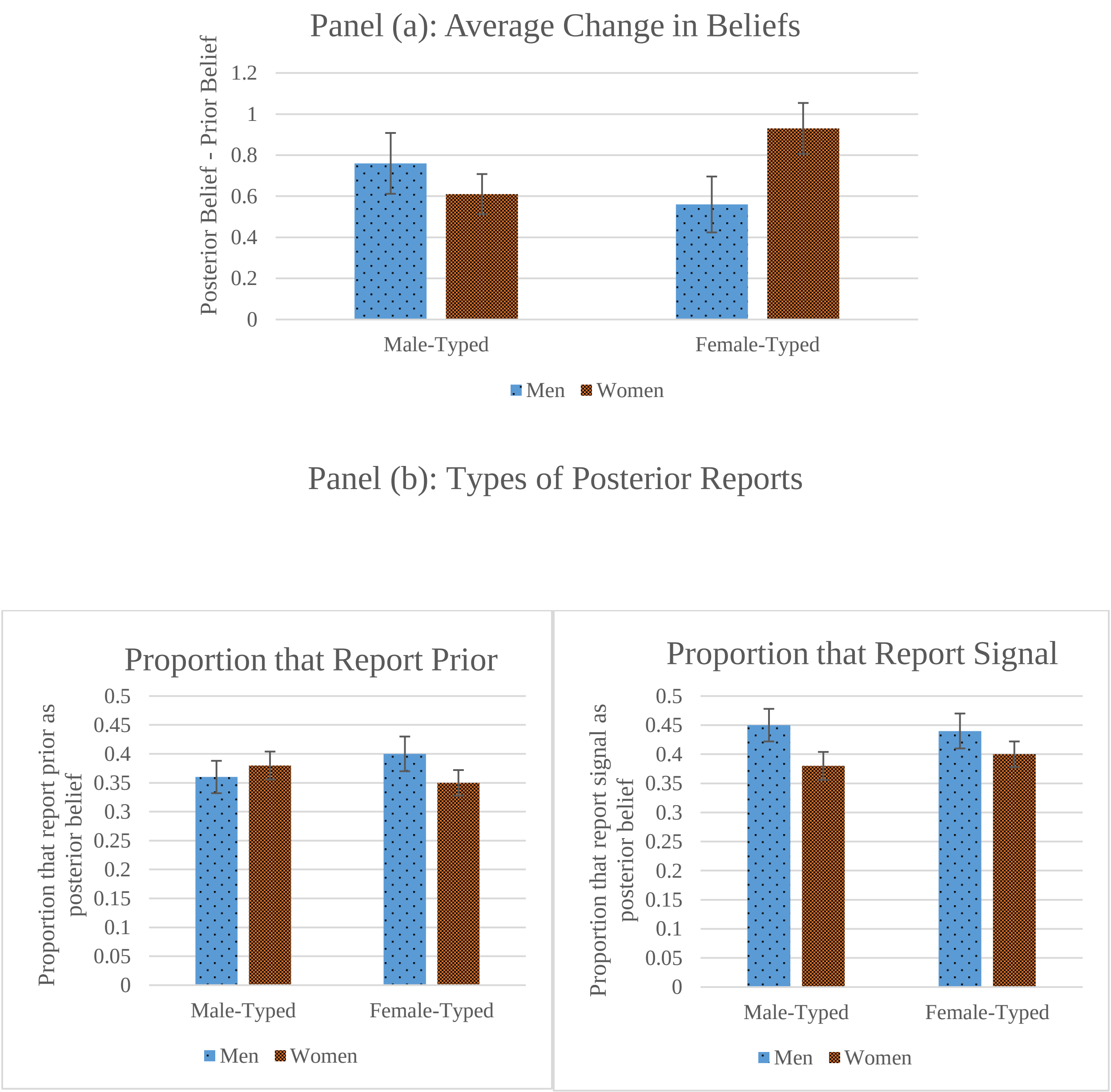
In Appendix Table B5, we show that Results 1 and 1a are robust to more flexible approaches for controlling for performance. We modify the specifications of Table I by removing our linear control for score and add a full set of fixed effects for score. Under this approach, the effect of self-stereotyping is reduced by approximately ¼, falling from 0.29 to 0.22 for prior beliefs and from 0.24 to 0.16 in posterior beliefs. These effects remain significant at the p<0.01. More importantly, our main message remains unchanged: across both treatments, gender stereotypes continue to predict beliefs after the provision of noisy feedback.

Self-Stereotyping in Posteriors is Not Well-Explained by the Bayesian Model

Is the remaining impact of stereotypes on posterior beliefs well-explained by the Bayesian model? Or, instead, do gender stereotypes predict systematic departures from the Bayesian benchmark? This requires understanding exactly how individuals update their beliefs in response to information. To provide a sense of how gender stereotypes shape belief updates, Figure III, Panel (a) graphs the average change in beliefs (posterior guess of score – prior guess of score). We split the data by the gender-type of the category and gender. We see that, on average, participants' beliefs of own score increase after seeing the information, becoming directionally more accurate on average. But, the extent of this updating varies. In the male-typed categories, men's beliefs increase directionally more than women's (p=0.10, estimated from a regression of type of report on gender that clusters standard errors at individual level), while in the female-typed categories, women's beliefs increase significantly more than men's (p<0.001). Visually, we can see that the

[14] Again, the results look quite similar if we instead predict the mean of the posterior belief given the reported distributions over possible scores, or if we use slider scale perceptions instead of average gender gaps in performance to account for stereotypes, or predict beliefs of relative performance. See Appendix Table B4.

gender stereotype of the domain seems to matter for how individuals adjust their beliefs in response to new information.


Panel (a): Average Change in Beliefs

Panel (b): Types of Posterior Reports

Proportion that Report Prior

Proportion that Report Signal

Figure III. Overview of Evidence on Posterior Beliefs

Error bars illustrate 95% confidence intervals. Male-typed are categories that have an average positive value on slider scale rating (Cars, Sports, Videogames, Business); female-typed are categories that have an average negative value on slider scale rating (Kardashians, Disney, Art, Verbal).

In Panel (b), we focus on the "type" of posterior report. In line with the Bayesian model, the two most frequent types of reports are (i) participants reporting their prior guess as their posterior guess, and (ii) participants reporting the signal as their posterior guess. Essentially, participants see the feedback and either stick to what they had believed initially, or, they update fully (reporting the signal as their posterior). The figure reveals that the proportion of each type of report varies by gender and gender-type of category. In male-typed categories, women are directionally more likely to stick to their prior beliefs than men (p=0.22); this pattern reverses in female-typed categories (p<0.01 on gender difference within female-typed categories, p<0.01 on the difference-in-difference). Men are more likely to report the signal as their posterior than women are in both male-typed categories (p<0.01) and female-typed categories (p<0.05), but this gender gap is directionally larger in male-typed categories (p=0.24 on the difference-in-difference).

The question we want to ask is whether this differential updating by gender-type of the category is well-explained by the Bayesian model. To answer this question, we estimate Equation (3), predicting a participant's observed posterior belief from the Bayesian prediction for her posterior. Conditional on these Bayesian predictions, is there a remaining impact of own gender advantage? If the updating we observe is in line with the Bayesian model, we should see no significant residual impact of gender or stereotypes once we include the Bayesian prediction in the model. Table II presents the results. For convenience, Column I reproduces Column II from Table I. This provides a useful reference point, as we can compare the estimated coefficients on own gender advantage with and without controlling for the Bayesian prediction.

While the Bayesian prediction does have significant predictive power for observed posteriors, we continue to estimate a significant role for both gender and stereotypes *on top of* the role that the Bayesian model predicts. For every 1-point increase in own gender advantage, we estimate that beliefs of own performance, conditional on the Bayesian prediction, increase by 0.15 points. Stereotypes bias the way participants update beliefs in response to information.[15] Note that, again, this result is robust to using score fixed effects to control for performance (see Appendix Table B7). We also estimate a significant male-female gap in posterior beliefs averaging across the categories, conditional on the Bayesian prediction. We find that women report posteriors roughly 0.3 points worse than men, on average, conditional on having the same Bayesian-predicted posterior.

Result 2: Gender stereotypes predict posterior beliefs, in excess of what the Bayesian model would predict.

[15] In Appendix Table B6, we show that Result 2 is robust to other approaches. We repeat this analysis but using the mean of the Bayesian-predicted posterior and the mean of the posterior belief distribution and find very similar results. We also repeat this analysis but using the Bayesian prediction from a smoothed prior where each possible score receives non-zero prior probability. Again, the results are very similar.

We now turn our attention to unpacking these systematic deviations from the Bayesian benchmark, with a particular focus on conservatism and asymmetry.

Table II. Documenting Systematic Deviations from the Bayesian Benchmark

	OLS Predicting Belief of Score $Belief_{i,j,t}$	
	Posterior Beliefs	
	I	II
$Female_i$	-0.33***	-0.29***
	(0.092)	(0.083)
Own Gender Advantage $\left(\overline{Score}_{j,G} - \overline{Score}_{j,-G}\right)$	0.24***	0.15***
	(0.022)	(0.021)
$Score_{i,j}$	0.79***	0.39***
	(0.012)	(0.017)
$Bayes_{i,j}$		0.45***
		(0.016)
Controls	Yes	Yes
Adjusted R-squared	0.618	0.681
Clusters (Obs.)	1989 (5967)	1989 (5967)

Notes: * indicates significance at p<0.10, ** at p<0.05, *** at p<0.01, and **** at p<0.001. Controls are a dummy for whether participant attended high school in the United States, dummies for each race category, dummies for each education category, fixed effects for round, average score in the category for the participant's own gender, and number of ASVAB questions answered correctly.

Gender Stereotypes and Responsiveness of Beliefs

We move to the model from Equation (4), predicting an individual's observed posterior belief from only her Bayesian predicted posterior. The estimated coefficient on the Bayesian prediction is a measure of responsiveness, with values greater than one reflecting more responsiveness of beliefs than the Bayesian model would predict and values less than one reflecting conservatism in responses to information relative to the Bayesian benchmark. We de-mean the Bayesian predictions to provide more intuitive interpretations of the constants. Table III presents the results, splitting the sample by gender and gender-type of the domain.

Table III. Responsiveness of Belief Updating by Gender and Gender Stereotypes

	OLS Predicting Belief of Score $Belief_{i,j,t}$			
	Posterior Beliefs			
	Male-Typed Domains		Female-Typed Domains	
	Men	Women	Men	Women
$Bayes_{i,j}$ De-meaned	0.78***	0.61***	0.70***	0.81***
	(0.021)	(0.023)	(0.030)	(0.015)
Constant	7.11***	6.19***	6.78***	6.90***
	(0.080)	(0.078)	(0.090)	(0.061)
Adjusted R-squared	0.634	0.460	0.523	0.722
Cluster (Obs.)	735 (1213)	1097 (1779)	716 (1139)	1123 (1836)

Notes: * indicates significance at p<0.10, ** at p<0.05, *** at p<0.01, and **** at p<0.001. Male-typed are categories that have an average positive value on slider scale rating (Cars, Sports, Videogames, Business); female-typed are categories that have an average negative value on slider scale rating (Kardashians, Disney, Art, Verbal). The Bayesian predictions are de-meaned by differencing out the mean Bayesian prediction in the full sample.

Overall, we replicate past findings of conservatism (or, in the language of Benjamin (2019), under-inference from the signals). For both men and women, across both gender congruent and gender incongruent categories, we estimate a slope on the Bayesian prediction significantly less than 1. Of course, our main interest lies in whether the extent of conservatism varies with the gender-type of the domain. We find that, for both men and women, posteriors are more responsive to the Bayesian prediction when the domain is gender congruent. Men's responsiveness falls from 0.78 to 0.70 when we move from male-typed to female-typed categories; women's responsiveness falls from 0.81 to 0.61 when moving from female-typed to male-typed categories. Notice that the adjusted R-squared of the model also drops considerably when moving from gender congruent to gender incongruent domains. Clearly, the Bayesian model better predicts posterior beliefs in gender congruent categories.[16]

Result 3: Men and women update their beliefs less conservatively in gender congruent domains.

[16] This result is unchanged if we use smoothed prior belief distributions to generate the Bayesian predictions. See Appendix Table B8.

Past literature has suggested that individuals are more conservative (under-infer more from signals, are less responsive relative to the Bayesian model) in more ego-relevant domains (for instance, forming beliefs about their own IQ compared to updating beliefs about the number of red balls in an urn). A natural question is whether this ego-relevance finding could help to explain our results. For instance, one could hypothesize that more gender congruent categories are also more ego-relevant for individuals on average. This would lead us to expect individuals to be more conservative, or under-infer more, in these more gender congruent categories, because they are more ego-relevant. However, we find that individuals are *more* responsive and better explained by the Bayesian model in more gender congruent categories, the opposite pattern. We discuss this more in our conclusion.

Good News, Bad News

Finally, we can consider asymmetry in updating in response to "good" or "bad" news, depending upon the gender-type of the domain. Recall that we define good news as receiving a signal greater than or equal to true score, and bad news as a signal less than true score. Assignment to good or bad news, defined this way, is random conditional on score and prior beliefs.

Figure IV presents the average change in beliefs, split by gender, gender-type of the domain, and type of news received. We observe that both men and women revise their beliefs upward on average in response to good news, but the magnitudes of these adjustments depend on the gender-type of the domain. Within male-typed domains, men change their beliefs in response to good news directionally more than women (p=0.20). In female-typed domains, this pattern is reversed: women change their beliefs significantly more than men in response to good news in female-typed domains (p<0.01 for gender difference within domain, p<0.01 on the difference-in-difference across domain types). Similarly, reactions to bad news also seem to depend on the gender-type of the domain, though there are no statistically significant differences. Within male-typed domains, women adjust their beliefs downward in response to bad news more so than men on average, while in female-typed domains, it is men who adjust their beliefs downward more in response to bad news.

Turning to regression analysis, we augment the model of Table III by adding an indicator for having received good news and the interaction of this indicator with the de-meaned Bayesian prediction, implementing Equation (5). We split the data according to gender-type of the domain and gender. The results are presented in Table IV, Panel A.

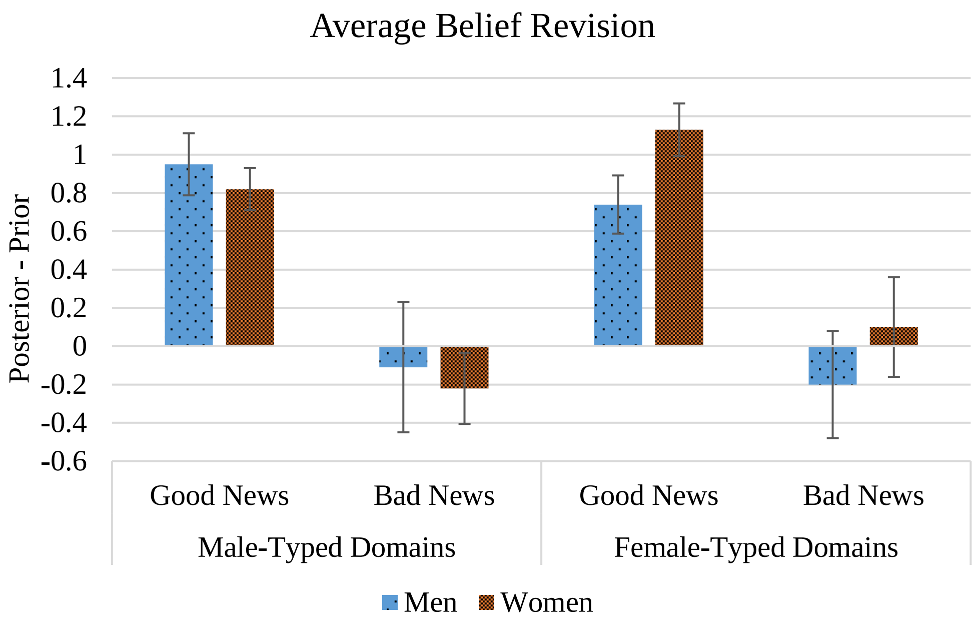


Figure IV. Average Belief Revisions after Good and Bad News

Error bars illustrate 95% confidence intervals. Male-typed are categories that have an average positive value on slider scale rating (Cars, Sports, Videogames, Business); female-typed are categories that have an average negative value on slider scale rating (Kardashians, Disney, Art, Verbal).

Let's start by considering male-typed domains. The constants in Columns I and II tell us that, on average, men report more optimistic beliefs than women after bad news. This is also the case after good news. Furthermore, we find no evidence of "generalized optimism" for men nor women, as beliefs after good news are shifted down, not up, relative to what the Bayesian model would predict for each.[17] Turning to responsiveness, we estimate that men and women display similar responsiveness to bad news in male-typed domains (with an estimated coefficient on the Bayesian prediction of approximately 0.75 for both men and women). However, stark differences emerge when considering responsiveness to good news. Men are estimated to be significantly more responsive to good news than women: estimated coefficient of 0.79 for men and 0.59 for women (p<0.01, estimated using the interacted model presented in Appendix Table B9).

[17] It seems likely that this finding is tied to the baseline level of under-confidence in our data. Because individuals are under-estimate their scores in their priors and are conservative relative to the Bayesian benchmark, this can combine to produce "generalized pessimism" after good news. In contrast, previous studies that have considered environments with baseline over-confidence could more readily generate "generalized optimism" after good news through conservatism.

Table IV. Good News and Bad News

Panel A: Regression Output

	OLS Predicting Belief of Score $Belief_{i,j,t}$			
	Posterior Beliefs			
	Male-Typed Domains		Female-Typed Domains	
	Men	Women	Men	Women
$Bayes_{i,j}$	0.76***	0.75***	0.79***	0.92***
	(0.052)	(0.040)	(0.055)	(0.022)
$GoodNews_{i,j}$	-0.34*	-0.44***	0.017	-0.39***
	(0.20)	(0.15)	(0.18)	(0.14)
$Bayes_{i,j}\ x\ GoodNews_{i,j}$	0.030	-0.16***	-0.11*	-0.12***
	(0.057)	(0.048)	(0.066)	(0.028)
Constant	7.38***	6.59***	6.79***	7.24***
	(0.18)	(0.12)	(0.15)	(0.12)
Adjusted R-squared	0.635	0.464	0.524	0.725
Cluster (Obs.)	735 (1213)	1097 (1779)	716 (1139)	1123 (1836)

Notes: * indicates significance at p<0.10, ** at p<0.05, *** at p<0.01, and **** at p<0.001. Good news is a dummy that takes 1 if the signal received was greater than or equal to true score. Male-typed are categories that have an average positive value on slider scale rating (Cars, Sports, Videogames, Business); female-typed are categories that have an average negative value on slider scale rating (Kardashians, Disney, Art, Verbal). We de-mean the Bayesian predictions by subtracting out the mean Bayesian prediction for the sample.

Panel B: Estimated Responsiveness by News Type and Gender Congruence

	Gender Congruent Domain	**Gender Incongruent Domain**
Good News	MEN: 0.79 WOMEN: 0.80	MEN: 0.68 WOMEN: 0.59
Bad News	MEN: 0.76 WOMEN: 0.92	MEN: 0.79 WOMEN: 0.75

These patterns look quite different when we turn to female-typed domains. While we continue to find no evidence of generalized optimism, many of the gender differences we observed for male-typed domains are reversed. In female-typed domains, women report more optimistic beliefs than men after receiving bad news. And, we find that women are more responsive – both to good news and bad – than men. For bad news, we estimate a responsiveness of 0.92 for women and 0.79 for men, p<0.01. After receiving good news, women are also more responsive than men, reversing the pattern we saw for male-typed domains (estimated coefficients of 0.80 and 0.68 for women and men, respectively, p<0.01).

Result 4a: The gender gap in responsiveness to good news depends upon the gender-type of the domain. In male-typed domains, men are more responsive to good news than women. But, in female-typed domains, the gap is reversed.

(Null) Result 4b: The gender gap in responsiveness to bad news does not vary systematically with the gender-type of the domain.

Panel B organizes the results from Panel A by gender congruence to help pull together the findings. To summarize, gender stereotypes play an important role in predicting reactions to good and bad news. Men and women are similarly responsive to good news when it comes in gender congruent domains: a responsiveness of approximately 0.80 for both. But, both men and women are less responsive to that same good news when it instead comes in a gender incongruent domain: men's responsiveness falls to 0.68 in female-typed domains, and women's falls to 0.59 in male-typed domains. It is as if individuals react more to "stereotype-confirming" information – good news in a gender congruent task – than "stereotype-disconfirming" information – good news in a gender incongruent task. Stereotypes shape how responsive individuals are to good news.

In Appendix Table B10, we replicate these results using the Bayesian predictions that are produced by smoothed priors. We obtain very similar results. And, in Appendix Table B11, we replicate these results while defining good and bad news relative to priors, rather than relative to true scores. While this introduces potentially important selection into the receipt of good and bad news, it likely corresponds more closely to participants' impressions of whether news is good or bad. We find similar results.

V. ROBUSTNESS OF RESULTS UNDER OTHER APPROACHES

In this section, we briefly discuss the robustness of our results to alternative approaches.

The Grether Model

While our econometric framework is useful for highlighting the role of gender stereotypes, while controlling for important potential confounds such as individual performance, gender-specific performance, and demographic controls, it is a departure from one popular method of analyzing deviations from the Bayesian model. In his comprehensive handbook chapter, Benjamin (2019) organizes his analysis around the Grether (1980) model that decomposes deviations from the Bayesian model into biased use of conditional likelihoods and biased use of prior beliefs. In Appendix Section E, we re-visit our analysis through the lens of this model, offering reinforcement of our main findings and deeper connection with past related work. We briefly summarize those findings here.

Using the Grether model, we observe significant under-inference from signals among our participants on average, and significantly greater under-inference in gender incongruent domains. That is, individuals treat signals as significantly less informative than they truly are, particularly in gender incongruent domains. While we also observe biased use of prior beliefs (what Benjamin refers to as base-rate neglect), under-inference is the relatively larger deviation from Bayesian updating in our setting. Again, we estimate that individuals are more responsive to good news (under-infer less from the signal) when it arrives in a gender congruent domain than when it arrives in a gender incongruent domain.

This model also allows us to consider how our results relate to other belief updating biases that have been observed in previous work, such as confirmation bias. This is the concept that individuals under-infer less from signals that confirm their priors. Our results are not well-explained by confirmation bias. In fact, in our setting, nearly all of our signals are disconfirming, as defined by Benjamin (2019) and Charness and Dave (2017). The rate of disconfirming signals is indistinguishable across gender incongruent and congruent domains, at approximately 86% in each. Thus, the fact that individuals under-infer less from signals in more congruent domains (and provide posteriors closer to Bayesian predictions in congruent domains) is not driven by the fact that they receive a larger share of confirming signals in those domains. We work through this formally in Appendix Section E.

Beyond the Bayesian Model

Our analysis asked, given the same Bayesian prediction for behavior, do men and women vary in their posterior beliefs depending upon the gender-type of the domain. This allowed us to conclude that the impact of gender stereotypes on posterior beliefs we observed marked a systematic deviation from the Bayesian benchmark. And, we documented the nature of these deviations.

But, our data also enables us to ask a different question. Are there other observable features of participant priors or situational characteristics that help to explain the impact of stereotyping that we observe? While our previous analysis highlighted the extent of non-Bayesian updating, this more demanding analysis explores whether the kitchen sink of observables that we have can help to explain the non-Bayesian impact of stereotypes. This question is not a typical focus of work on belief updating. However, assessing it allows us to ask whether there are characteristics of prior beliefs or other features that correlate with stereotype-driven updating.

From a theoretical perspective, the Bayesian prediction incorporates all information about the participant prior that should be relevant for the posterior reported. In practice, however, different priors, despite generating the same Bayesian prediction, might intuitively produce different expected posteriors. For example, consider two participants both with a true score of 9 in a category. Suppose Participant A reports a prior guess of "7", while Participant B reports a prior guess of "9". Both then receive a signal of "9". It may be the case that the Bayesian prediction for both participants is to report the signal as their posterior guess (assuming Participant A puts sufficient weight on the possibility of her score being 9 in her prior distribution). However, despite the Bayesian model making the same prediction, we may expect different responses – one could reasonably predict that Participant B is much more likely to report 9 as her posterior guess than Participant A would be, as the signal is more in line with Participant B's prior belief.

We report the results in Appendix Table B12, appending Table II with an additional column (Column III) that includes a battery of other characteristics of the prior. Including the reported mode of the prior in addition to the Bayesian prediction adds a lot of explanatory power to the model. Similarly, participants who have more variant priors and are more positively-skewed on average report greater posteriors conditional on other observables. The hypothesis that some priors make it "easier" to follow the Bayesian prediction seems to be supported by the data. Conditioning on this additional information produces a smaller (approximately 1/3 the size), but still significant, role for stereotypes. This suggests that, indeed, there are characteristics of prior beliefs that help to explain the stereotype-driven deviations from the Bayesian model that we predict. The degree of initial under-confidence plays a particularly critical role.

In Column IV, we take the analysis a step further, replacing linear controls for participant score and prior beliefs with a full set of fixed effects for each. In this model, the estimated impact of stereotypes is no longer statistically significant, falling to 0.02 (p=0.23). This analysis helps us to caveat our conclusions. Stereotypes produce deviations from the Bayesian model, with participants adjusting their beliefs more in response to feedback in more gender congruent domains. However, a substantial portion of this non-Bayesian behavior does seem to be explained by characteristics of participant priors. The significant under-confidence we observe in gender incongruent domains seems to drive the under-reaction to information, albeit to an extent that is not well-explained by the Bayesian model.

VI. CONCLUSION

There is increasing evidence that stereotyped beliefs influence important economic decisions – such as willingness to answer when unsure (Baldiga 2014, Coffman and Klinowski 2020), willingness to contribute ideas (Coffman 2014, Chen and Houser 2017, Bordalo et al 2019), willingness to compete (Niederle and Vesterlund 2007), and willingness to lead (Born et al 2020). Given this growing consensus, an important question is how persistent are these biased beliefs and how do they evolve over time. In this paper, we take an important step toward addressing this question, asking how individuals respond to feedback about their abilities across different domains.

We find a significant role for self-stereotyping in predicting beliefs of objective ability absent feedback. We then show that self-stereotyping is also highly predictive of beliefs after noisy feedback. Within our setting, more informative feedback is no more effective in reducing reliance on gender stereotypes. The impact of gender stereotypes on posterior beliefs operates, in part, through biases in updating. Holding fixed the Bayesian prediction for beliefs, individuals hold more optimistic beliefs about their performance as the domain becomes more gender congruent.

Both men's and women's beliefs are better predicted by the Bayesian model in gender congruent categories. In gender incongruent categories, participants' posterior beliefs are stickier to prior beliefs, and less responsive to good news in particular.

One concern is whether the results we find for these specific female and male-typed domains are likely to generalize to a variety of educational and professional contexts beyond this study. In a first step toward addressing this important consideration, we explore whether our results are weaker if we restrict to domains with likely greater external relevance: business, verbal skills, and art and literature. Appendix Table B13 presents the results. We find that the predictive power of stereotypes for prior and posterior beliefs is no weaker in this restricted sample. This, along with our evidence from our complementary study in Appendix A that explores beliefs about cognitive skills, shows that these patterns persist for more intellectually serious domains.

Our work advances our understanding of the ways in which individuals deviate from the Bayesian model in updating their beliefs. We see that individuals under-infer more from signals, particularly good news signals, in gender incongruent domains. These differences by gender-type of the domain do not seem well-explained by previously documented biases, such as confirmation bias. Past work has also shown that individuals are more conservative (relative to the Bayesian model) for more ego-relevant tasks. Thus, the extent to which more gender congruent categories are more "ego-relevant" for participants would push our results in the opposite direction of what we find.

While reconciling the mixed results in the literature is beyond the scope of this paper, a few possibilities seem worth noting. It could be the case that ego-relevance is not the right framework for thinking about how gender stereotypes shape belief formation. Or, or in addition, it could be the case that past findings of greater conservatism in more ego-relevant categories depend on the fact that, in those settings, individuals' prior beliefs are overconfident on average: when asked to adjust beliefs *downwards* on average, beliefs are stickier to priors in more ego-relevant domains. In our setting, however, individuals are underconfident on average in their priors, and are typically asked to adjust upwards. In this environment, more responsiveness, not less, in more ego-relevant categories may be the more reasonable prediction. This would put our finding of greater responsiveness – particularly in response to good news - in more gender congruent categories more in line with past work on the role of ego-relevance. Future work should delve more fully into how baseline levels of over and underconfidence contribute to these patterns. Indeed, our work seems to suggest that two main themes of past work in this area (good news - bad news asymmetries, and more conservatism in more ego-relevant domains) seem to depend in part on whether participant priors are over or underconfident on average.

We hypothesized that individuals would be more responsive to "stereotype-consistent" news: good news in a gender congruent domain, bad news in a gender incongruent domain. While we find significant evidence in favor of the former – more responsiveness to good news in gender congruent domains – we, perhaps surprisingly, do not find evidence of the latter. It may be the case that the significant degree of underconfidence in priors contributes to this pattern as well. We may have more measurement error in analyzing bad news, as relatively few of our participants receive a signal that might truly be considered bad news (a signal less than their prior belief). Those that do receive "truly" bad news are a highly-selected sample, individuals who were most overconfident in their prior beliefs. Future work should probe the case for increased responsiveness to stereotype-consistent news – both good and bad - across an array of settings, including those that are characterized by average overconfidence and average underconfidence.

More work is also needed to understand the mechanisms behind stereotype-driven updating. Recent work in cognitive economics offers some promising explanations why individuals might be more responsive to stereotype-consistent news. Drawing on insights from the memory literature, Bordalo et al (2022) lay out a model for how simulation based upon similar past experiences helps decision-makers to form beliefs. We can apply this idea in our setting. Consider a woman who receives good news in a domain; she asks herself, is this really my true score? She draws upon past experiences; similar experiences, even if objectively irrelevant, inform her beliefs, particularly if they come to mind easily. She might think about positive feedback she has received in the past in this area, or whether she can think of people like her – perhaps other women – who are knowledgeable in this domain. If she has received more positive feedback in female-typed domains compared to male-typed domains, or more easily recalls expert women in female-typed domains than male-typed domains, it is easier for her to simulate – or imagine – this good news reflecting her true performance. In this way, the simulation channel might lead individuals to be more likely to "believe" and respond to stereotype-consistent news. Future work should probe this potential mechanism.

Our results have potentially important implications for policy-makers, educators, and organizational leaders looking to address gender gaps in self-confidence, particularly in male-typed domains. While a natural policy suggestion for addressing under-confidence among talented women in male-typed domains is providing feedback about own ability, our results suggest that stereotypes may inhibit the effectiveness of this strategy. In our setting, convincing an individual of their talent in a gender incongruent domain is more difficult than convincing an individual of their talent in a gender congruent domain. This speaks to the persistence of self-stereotyping. Stereotypes do not just impact beliefs about ability when information is scarce; rather, it appears stereotypes also influence the way information is incorporated into beliefs, perpetuating initial biases.

REFERENCES

Baldiga, Katherine. 2014. "Gender Differences in Willingness to Guess." *Management Science* 60 (2): 434–48. https://doi.org/10.1287/mnsc.2013.1776.

Barber, Brad M., and Terrance Odean. 2001. "Boys Will Be Boys: Gender, Overconfidence, and Common Stock Investment." *The Quarterly Journal of Economics* 116 (1): 261–92. https://doi.org/10.1162/003355301556400.

Barron, Kai. 2021. "Belief Updating: Does the 'Good-News, Bad-News' Asymmetry Extend to Purely Financial Domains?" *Experimental Economics* 24 (1): 31–58. https://doi.org/10/ghrgx5.

Bénabou, Roland, and Jean Tirole. 2002. "Self-Confidence and Personal Motivation." *The Quarterly Journal of Economics* 117 (3): 871–915. https://doi.org/10.1162/003355302760193913.

Benjamin, Daniel. 2019. Errors in probabilistic reasoning and judgment biases. *Handbook of Behavioral Economics: Applications and Foundations 1, 2*, 69-186.

Beyer, Sylvia. 1990. "Gender Differences in the Accuracy of Self-Evaluations of Performance." *Journal of Personality and Social Psychology* 59 (5): 960–70.

———. 1998. "Gender Differences in Self-Perception and Negative Recall Biases." *Sex Roles* 38 (1): 103–33. https://doi.org/10.1023/A:1018768729602.

Beyer, Sylvia, and Edward M. Bowden. 1997. "Gender Differences in Self-Perceptions: Convergent Evidence from Three Measures of Accuracy and Bias." *Personality and Social Psychology Bulletin* 23 (2): 157–72. https://doi.org/10.1177/0146167297232005.

Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer. 2016. "Stereotypes." *The Quarterly Journal of Economics* 131 (4): 1753–94. https://doi.org/10.1093/qje/qjw029.

Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer. 2019. "Beliefs about Gender." *American Economic Review* 109 (3): 739–73. https://doi.org/10/gfxjck.

Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer. 2022. Imagining the Future: Memory, Simulation, and Beliefs about Covid. *Working Paper*.

Born, Andreas, Eva Ranehill, and Anna Sandberg. 2022. "Gender and Willingness to Lead: Does the Gender Composition of Teams Matter?" *The Review of Economics and Statistics*, January, 1–17.

Buser, Thomas, Leonie Gerhards, and Joël van der Weele. 2018. "Responsiveness to Feedback as a Personal Trait." *Journal of Risk and Uncertainty* 56 (2): 165–92. https://doi.org/10.1007/s11166-018-9277-3.

Buser, Thomas, Muriel Niederle, and Hessel Oosterbeek. 2014. "Gender, Competitiveness, and Career Choices." *The Quarterly Journal of Economics* 129 (3): 1409–47. https://doi.org/10.1093/qje/qju009.

Campbell, W. Keith, and Constantine Sedikides. 1999. "Self-Threat Magnifies the Self-Serving Bias: A Meta-Analytic Integration." *Review of General Psychology* 3 (1): 23–43. https://doi.org/10.1037/1089-2680.3.1.23.

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

Chen, Jingnan, and Daniel Houser. 2017. “Gender Composition, Stereotype and the Contribution of Ideas by Jingnan Chen, Daniel Houser :: SSRN.” *Working Paper*, May. https://papers-ssrn-com.ezp-prod1.hul.harvard.edu/sol3/papers.cfm?abstract_id=2989049&download=yes.

Coffman, Katherine B. 2014. “Evidence on Self-Stereotyping and the Contribution of Ideas.” *The Quarterly Journal of Economics* 129 (4): 1625–60. https://doi.org/10.1093/qje/qju023.

Coffman, Katherine B, Manuela R. Collis, and Leena Kulkarni. 2023. “Whether to Apply.” Forthcoming.*Management Science*.

Coffman, Katherine B., and David Klinowski. 2020. “The Impact of Penalties for Wrong Answers on the Gender Gap in Test Scores.” *Proceedings of the National Academy of Sciences* 117 (16): 8794–8803. https://doi.org/10/gjnmd3.

Coutts, Alexander. 2019. “Good News and Bad News Are Still News: Experimental Evidence on Belief Updating.” *Experimental Economics* 22 (2): 369–95. https://doi.org/10/gjnmd7.

Deaux, Kay, and Elizabeth Farris. 1977. “Attributing Causes for One’s Own Performance: The Effects of Sex, Norms, and Outcome.” *Journal of Research in Personality* 11 (1): 59–72. https://doi.org/10.1016/0092-6566(77)90029-0.

Dreber, Anna, Emma von Essen, and Eva Ranehill. 2011. “Outrunning the Gender Gap—Boys and Girls Compete Equally.” *Experimental Economics* 14 (4): 567–82. https://doi.org/10.1007/s10683-011-9282-8.

Christoph Drobner, Sebastian J. Goerg. 2022. “Motivated Belief Updating and Rationalization of Information.” *Working Paper*.

Eil, David, and Justin M. Rao. 2011. “The Good News-Bad News Effect: Asymmetric Processing of Objective Information about Yourself.” *American Economic Journal: Microeconomics* 3 (2): 114–38.

Ertac, Seda. 2011. “Does Self-Relevance Affect Information Processing? Experimental Evidence on the Response to Performance and Non-Performance Feedback.” *Journal of Economic Behavior & Organization* 80 (3): 532–45. https://doi.org/10.1016/j.jebo.2011.05.012.

Exley, Christine L., and Judd B. Kessler. "The gender gap in self-promotion." *The Quarterly Journal of Economics* 137, no. 3 (2022): 1345-1381. https://doi.org/10.1093/qje/qjac003.

Gillen, Ben, Erik Snowberg, and Leeat Yariv. "Experimenting with measurement error: Techniques with applications to the caltech cohort study." *Journal of Political Economy* 127, no. 4 (2019): 1826-1863.

Gotthard-Real, Alexander. 2017. “Desirability and Information Processing: An Experimental Study.” *Economics Letters* 152 (March): 96–99. https://doi.org/10.1016/j.econlet.2017.01.012.

Grether, D.M., 1980. "Bayes rule as a descriptive model: the representativeness heuristic." *The Quarterly Journal of Economics* 95 (3), 537–557.

Große, Niels Daniel, and Gerhard Riener. 2010. “Explaining Gender Differences in Competitiveness: Gender-Task Stereotypes.” Working Paper 2010,017. Jena Economic Research Papers. https://www.econstor.eu/handle/10419/32599.

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

Heider, Fritz. 1958. *The Psychology of Interpersonal Relations*. Psychology Press.

Kőszegi, Botond, and Matthew Rabin. 2006. “A Model of Reference-Dependent Preferences.” *The Quarterly Journal of Economics* 121 (4): 1133–65.

Lichtenstein, Sarah, and Baruch Fischhoff. 1977. "Do Those Who Know More Also Know More about How Much They Know? *Organizational Behavior and Human Performance 20:* 159-183.

Lundeberg, Mary A., Paul W. Fox, and Judith Punćcohaŕ. 1994. “Highly Confident but Wrong: Gender Differences and Similarities in Confidence Judgments.” *Journal of Educational Psychology* 86 (1): 114–21. http://dx.doi.org/10.1037/0022-0663.86.1.114.

Lusardi, Annamaria, Olivia S. Mitchell, and Vilsa Curto. 2010. “Financial Literacy among the Young.” *The Journal of Consumer Affairs* 44 (2): 358–80.

Mobius, Markus M., Muriel Niederle, Paul Niehaus, and Tanya S. Rosenblat. 2022. “Managing Self-Confidence: Theory and Experimental Evidence.” *Management Science.*

Niederle, Muriel, and Lise Vesterlund. 2007. “Do Women Shy Away From Competition? Do Men Compete Too Much?” *The Quarterly Journal of Economics* 122 (3): 1067–1101. https://doi.org/10.1162/qjec.122.3.1067.

Pan, Siqi. 2019. “The Instability of Matching with Overconfident Agents.” *Games and Economic Behavior* 113: 396–415. https://doi.org/10/gjnmfg.

Pulford, Briony D., and Andrew M. Colman. 1997. “Overconfidence: Feedback and Item Difficulty Effects.” *Personality and Individual Differences* 23 (1): 125–33. https://doi.org/10.1016/S0191-8869(97)00028-7.

Rabin, Matthew, and Joel L. Schrag. 1999. “First Impressions Matter: A Model of Confirmatory Bias.” *The Quarterly Journal of Economics* 114 (1): 37–82. https://doi.org/10.1162/003355399555945.

Reuben, Ernesto, Paola Sapienza, and Luigi Zingales. 2014. “How Stereotypes Impair Women’s Careers in Science.” *Proceedings of the National Academy of Sciences*, March. https://doi.org/10.1073/pnas.1314788111.

Schwardmann, Peter, and Joël van der Weele. 2019. “Deception and Self-Deception.” *Nature Human Behaviour* 3 (10): 1055–61. https://doi.org/10/gf7qqm.

Shastry, Gauri Kartini, Olga Shurchkov, and Lingjun Lotus Xia. 2020. “Luck or Skill: How Women and Men React to Noisy Feedback.” *Journal of Behavioral and Experimental Economics* 88 (October): 101592. https://doi.org/10/ghmtmm.

Shurchkov, Olga. 2012. “Under Pressure: Gender Differences in Output Quality and Quantity under Competition and Time Constraints.” *Journal of the European Economic Association* 10 (5): 1189–1213. https://doi.org/10.1111/j.1542-4774.2012.01084.x.

Zimmermann, Florian. 2020. "The Dynamics of Motivated Beliefs." *The American Economic Review* 110(2), 337-61.

APPENDIX A. MOTIVATING EVIDENCE IN THE COGNITIVE SKILLS DOMAIN

Before conducting the main experiment reported in the paper, we collected data from a simpler study that considered a single domain. We report those results below. The findings are largely consistent with the findings in our main experiment for male-typed domains.

Design

In this first study, participants take a test consisting of multiple-choice questions from the Armed Services Vocational Aptitude Battery (ASVAB). The ASVAB is an enlistment exam administered by the United States Armed Forces and taken annually by more than one million people (http://official-asvab.com/). In social science research, performance on the ASVAB has been used as a proxy for cognitive ability (see, for instance, Lusardi, Mitchell, and Curto 2010). We selected 30 total questions from five domains tested on the ASVAB: General Science, Arithmetic Reasoning, Math Knowledge, Mechanical Comprehension, and Assembling Objects. Participants have five minutes to answer as many questions as they can, and receive $0.20 for each correct answer if this portion of the experiment is selected for payment. Incorrect answers and skipped questions are not penalized.

Following their completion of the test, we elicit beliefs from participants. First, we ask each participant to guess their score -- their total number of correct answers -- on the test. We refer to this as a participant's prior belief of her absolute performance. Next, we ask each participant to provide a belief of relative ability. We ask them to consider how their performance on the test compared to the performance of all other participants completing the experiment. We asked them to choose which bucket they believed their relative performance fell into: 0 – 5th percentile, 5th – 20th percentile, 20th – 40th percentile, 40th – 60th percentile, 60th – 80th percentile, 80th – 95th percentile, 95th – 100th percentile. We explained these percentiles as identifying the percentage of other participants who performed better or worse than the participant. For each of these prior beliefs, we incentivize participants by offering them $0.10 if their guess is correct. In this way, we incentivize participants to provide the mode of their subjective belief.

Participants are then randomly assigned to one of two signal treatments, either the *60% Signal Treatment* or the *90% Signal Treatment.* Across both treatments, individuals receive a noisy signal of their performance on the test. With probability *p,* where *p* is either 0.6 or 0.9 depending on the treatment, the signal transmitted is exactly equal to their score on the test. With probability *1 – p,* the signal is equal to their score plus randomly-drawn "noise". The noise is drawn from a uniform distribution over non-zero integers between -5 and 5, that is: {-5, -4, -3, -2, -1, 1, 2, 3, 4, 5}.

After they see their signal, participants are asked to provide another guess of their score on the test, incentivized in the same way as the prior. We will refer to this belief as a participant's posterior belief of her absolute performance on the test.

Finally, we collect some minimal demographic information about the participant: her gender, whether she attended high school in the United States, her race, and her educational attainment. Note that this beliefs experiment was embedded within a larger experiment aimed at exploring individuals' decisions about when to apply for promotion opportunities. The remainder of the study also includes a second round of ASVAB problem-solving; this provides an additional measure of ASVAB performance, which we can use to address potential measurement error concerns. All interventions related to this larger study occur after the beliefs experiment (but before the demographic information is elicited). That experiment is described in detail in Coffman, Collis, and Kulkarni (2022).

The experiment was run on Amazon Mechanical Turk in May 2018 with a total of 1,502 workers, of which 981 are assigned to one of the two signal treatments (the remaining participants receive no signal and so are excluded from this analysis as their posterior beliefs are not elicited). The study was advertised as a 30-minute academic research study that guaranteed a completion payment of $2.50 with the possibility of additional incentive pay. The study was restricted to workers with a United States based IP address who had completed at least 100 tasks (called HITs) and had an approval rating by previous Amazon Mechanical Turk requesters of at least 95%. The study contains understanding questions and a participant must answer those understanding questions correctly in order to complete the study.

Results

In Appendix Table A1, we report summary statistics for our participants. There are significant gender differences in performance (and beliefs) on the ASVAB test, suggestive that this is a male-typed domain. We compute score as the total number of correct answers provided during the timed test. Men earn an average score of 11.3 (4.57 SD), while women earn an average score of 9.57 (4.20 SD). We reject the null of equality using a two-tailed t-test with $p<0.001$. Just as we see in our main study, on average, participants underestimate their absolute performance on the test when stating their prior beliefs. Men believe they answered 8.89 questions correctly on average (4.16 SD), while women believe they answered 7.26 questions correctly on average (3.86 SD) ($p<0.001$).

In Appendix Table A2, we parallel our empirical approach from the main text, regressing prior and then posterior beliefs of score on participant gender and performance. In the main study, we used a range of domains that varied in their gender-type. Here, there is a single, male-typed domain. Therefore, our key coefficient of interest is the coefficient on gender. We expect that women will be less confident in their performance than men, conditional on performance, and we ask whether feedback reduces that gap. One concern is that performance in Round 1 is a noisy measure of true ability, and unobserved ability may be correlated with gender. Following the recommendation of Gillen et al (2019), we can address this through the inclusion of multiple measures. In this study, we have both a Round 1 performance and a Round 2 performance, where the Round 2 quiz involves similar ASVAB problems (the correlation between Round 1 and Round 2 score is 0.51). Therefore, we include in our model both Round 1 and Round 2 performance.

Table A1. Summary Statistics for Cognitive Skills Study

	Men	**Women**	**P-value**
White	0.80	0.81	0.60
Black	0.06	0.10	0.04
Asian	0.10	0.06	0.03
Attended HS in US	0.98	0.97	0.49
HS Only	0.10	0.08	0.22
Some College/Assoc.	0.35	0.37	0.43
Bachelors	0.40	0.42	0.58
Advanced Degree	0.15	0.13	0.39
Treatment Assignment			
60% Signal	0.51	0.47	0.20
90% Signal	0.49	0.53	0.20
Mean Score (out of 30)	11.3	9.57	<0.001
N	518	463	

Column I produces estimates for prior beliefs; we can contrast these with the estimates for posterior beliefs in Column II. Conditional on performance, we estimate that women state prior beliefs 0.6 points lower than men (Column I, p<0.01).[18]

Column II presents the results for posterior beliefs. We see that women's beliefs remain about 0.45 points lower than men's after feedback. Overall, more men than women appear to be "convinced" by the signal. We find that 60% of men but only 49% of women report the signal they observe as their posterior belief (p<0.01), while 25% of men and 30% of women stick to their prior belief (p<0.10). On average, men revise their beliefs upward by 1.50 points, while women revise their beliefs upward by 1.15 points (p<0.10). These gender gaps are qualitatively similar to what we observe for male-typed domains in our main study.

[18] In terms of relative ability, women believe they place 7.6 percentage points worse in the ability distribution compared to equally able men (p<0.001). Interestingly, even when we control for a participant's prior belief of her absolute score, women believe they place worse in the relative distribution than men do. This suggests that the relative beliefs gap is driven not just by women believing they earned worse scores in absolute terms, but also by women believing others were more likely to earn better scores.

In Columns III – VI, we present the results separately by signal accuracy, with Columns III and IV displaying prior and posterior beliefs for the 60% treatment and Columns V and VI presenting the results for the 90% treatment. For the 60% treatment, we see that feedback reduces the gender gap by just over 1/3. However, there is still a sizable and statistically significant gender gap post-feedback (0.57, p<0.05, Column IV). The results from the 90% treatment are perhaps more surprising. Feedback fails to reduce the initial gender gap within this highly informative treatment. While the gender gap post-feedback is smaller in the 90% treatment than in the 60% treatment, once we account for the initial imbalance across treatment in the gender gap in prior beliefs, we see that the 90% signal treatment seems to be less, not more, effective at closing the gender gap. Note that in many ways, the 90% signal accuracy treatment represents a near "best-case scenario" for feedback: this feedback is highly accurate, immediate, and task-specific. And yet, it still fails to eliminate the gender gap.

Table A2. Prior and Posterior Beliefs for Cognitive Skills Task

	OLS Predicting Belief of Score $Belief_{i,t}$					
	Full Sample		60% Signal		90% Signal	
	I	II	III	IV	V	VI
	Prior Beliefs	Posterior Beliefs	Prior Beliefs	Posterior Beliefs	Prior Beliefs	Posterior Beliefs
$Female_i$	-0.60***	-0.45***	-0.87***	-0.57**	-0.30	-0.42**
	(0.20)	(0.15)	(0.29)	(0.24)	(0.28)	(0.20)
$Rd1Score_i$	0.61***	0.88***	0.59***	0.83***	0.64***	0.91***
	(0.025)	(0.020)	(0.038)	(0.032)	(0.035)	(0.025)
$Rd2Score_i$	-0.027	0.0096	-0.00063	0.010	-0.057	0.011
	(0.030)	(0.023)	(0.044)	(0.037)	(0.042)	(0.030)
Demographic Controls	Yes	Yes	Yes	Yes	Yes	Yes
Observations	981	981	481	481	500	500
Adjusted R-squared	0.471	0.750	0.455	0.688	0.484	0.801

Notes: * indicates significance at p<0.10, ** at p<0.05, *** at p<0.01, and **** at p<0.001. Demographic controls are a dummy for whether participant attended high school in the United States, dummies for each race category, and dummies for each education category.

For our main study, we asked whether the impact of gender stereotypes after feedback could be explained by the Bayesian model. In this simpler study, we cannot construct a Bayesian benchmark, as we do not elicit a full belief distribution for each individual. However, we can pursue a proxy: we can add prior beliefs and the signal received to our model from Table A2. We can ask how much of the residual gender gap can be explained by including this proxy for the Bayesian prediction. We present these results in Table A3. Columns I simply reproduces Column II from Table A2; Column II adds the participant's prior belief and the signal she received. We observe that. just as in our main study, proxies for the Bayesian model indeed have predictive power and explain some of the residual gender gap. The gender gap shrinks from 0.45 points to 0.25 points, remaining statistically significant (p<0.05).

Table A3. How Well Does a Bayesian-Model Proxy Explain Gender Gaps in Posteriors?

	OLS Predicting Belief of Score $Belief_{i,t}$	
	Posterior Beliefs	
	I	II
$Female_i$	-0.45***	-0.25**
	(0.15)	(0.12)
$Rd1Score_i$	0.88***	0.19***
	(0.020)	(0.038)
$Rd2Score_i$	0.0096	0.015
	(0.023)	(0.018)
$Prior\ Belief_i$		0.43***
		(0.019)
$Signal_i$		0.42***
		(0.032)
Demographic Controls	Yes	Yes
Observations	981	981
Adjusted R-squared	0.750	0.855

Notes: * indicates significance at p<0.10, ** at p<0.05, *** at p<0.01, and **** at p<0.001. Demographic controls are a dummy for whether participant attended high school in the United States, dummies for each race category, and dummies for each education category.

Together, our results from this simpler, single-domain paradigm suggest that the results we observe for the male-typed domains in our study persist in an environment with (i) an economically-meaningful domain, and (ii) signals that are more accurate on average.

APPENDIX B.

Table B1. Summary Statistics for Main Study

Panel A. Summary Statistics by Gender

	Men	Women	P-value from test of proportions
White	0.79	0.82	0.13
Black	0.08	0.09	0.43
Asian	0.08	0.05	<0.01
Attended HS in US	0.96	0.96	0.76
HS Only	0.10	0.09	0.41
Some College/Assoc.	0.33	0.39	<0.01
Bachelors	0.42	0.38	0.08
Advanced Degree	0.16	0.15	0.43
ASVAB Score (out of 5)ǂ	3.41	3.39	0.71
Treatment Assignment			
50% Signal	0.48	0.51	0.17
70% Signal	0.52	0.49	0.17
N	784	1,205	

ǂ Indicates p-value was from t-test comparing means, rather than test of proportions.

Panel B. Summary Statistics by Category

	Kard-ashians	Disney	Art	Verbal	Business	Video-games	Sports	Cars
Avg. Slider Scale Rating	-0.59	-0.43	-0.21	-0.21	0.17	0.44	0.48	0.52
Male Avg. Score	8.47	8.28	7.37	4.50	5.66	10.98	7.97	8.49
Female Avg. Score	10.63	11.47	7.71	4.16	4.73	8.15	6.16	7.51
Avg. Male Advantage	-2.17****	-3.19****	-0.34	0.33*	0.93****	2.83****	1.82****	0.98****
Male Avg. Prior Belief	5.58	6.57	6.16	5.47	5.46	8.99	6.65	6.37
Female Avg. Prior Belief	7.16	8.93	5.83	5.31	4.40	5.40	4.56	4.91
Avg. Male Advantage in Priors	-1.58****	-2.37****	0.34	0.16	1.06****	3.59****	2.09****	1.45****
Male Avg. Posterior Belief	6.68	7.39	6.72	5.25	5.72	10.15	7.29	7.35
Female Avg. Posterior Belief	8.78	10.34	6.73	5.08	4.46	6.39	5.11	5.73

Avg. Male Advantage in Posteriors	-2.10****	-2.95****	-0.01	0.17	1.25****	3.76****	2.18****	1.62****
N	740	746	748	741	741	738	762	751

Notes: * indicates significance at p<0.10, ** at p<0.05, *** at p<0.01, and **** at p<0.001 from a two-tailed t-test comparing average performance/beliefs of men and women.

Table B2. Robustness of Prior Beliefs to Other Approaches

OLS Predicting Belief $Belief_{i,j,t}$						
Prior Beliefs						
	Belief of Score	**Mean of Full Prior over Scores**	**Belief of Percentile Rank (1-100, 1 is Best)**	**Belief of Score**	**Mean of Full Prior over Scores**	**Belief of Percentile Rank (1-100, 1 is Best)**
	I	II	III	IV	V	VI
$Female_i$	-0.45***	-0.38***	4.62***	-0.48***	-0.41***	4.76***
	(0.099)	(0.10)	(0.87)	(0.099)	(0.10)	(0.87)
Own Gender Advantage $(\overline{Score}_{j,G} - \overline{Score}_{j,-G})$	0.29***	0.30***	-1.71***			
	(0.024)	(0.023)	(0.21)			
Own Gender Advantage in Perception				1.18***	1.22***	-7.07***
				(0.097)	(0.10)	(0.89)
$Score_{i,j}$	0.61***	0.62***	-2.20***	0.61***	0.62***	-2.20***
	(0.013)	(0.013)	(0.11)	(0.013)	(0.014)	(0.11)
Demographic Controls	Yes	Yes		Yes	Yes	
Adjusted R-squared	0.444	0.428	0.123	0.442	0.426	0.122
Clusters (Obs.)	1989 (5967)	1989 (5967)	1989 (5967)	1989 (5967)	1989 (5967)	1989 (5967)

Notes: * indicates significance at p<0.10, ** at p<0.05, *** at p<0.01, and **** at p<0.001. Demographic controls are a dummy for whether participant attended high school in the United States, dummies for each race category, dummies for each education category, fixed effects for round, average score in the category for the participant's own gender, and number of ASVAB questions answered correctly.

.

Table B3. Gender Differences and Stereotypes in Shapes of Priors

	OLS Predicting Prob. Assigned to Mode of Prior	OLS Predicting Range of Prior	OLS Predicting SD of Prior	OLS Predicting a Dummy if Prior Mean = Prior Median	OLS Predicting Right-Skewness of Prior (Mean – Median)
	I	II	III	IV	V
$Female_i$	1.46*	-0.62***	-0.15***	-0.030**	0.028*
	(0.85)	(0.18)	(0.050)	(0.013)	(0.016)
Own Gender Advantage $(\overline{Score}_{j,G} - \overline{Score}_{j,-G})$	0.53***	0.0024	0.0063	0.00069	-0.0055
	(0.18)	(0.033)	(0.0091)	(0.0030)	(0.0042)
$Score_{i,j}$	0.23**	-0.013	-0.011*	-0.0043***	0.00041
	(0.097)	(0.020)	(0.0055)	(0.0016)	(0.0022)
Demographic Controls	Yes	Yes	Yes	Yes	Yes
Adjusted R-squared	0.024	0.056	0.054	0.019	0.012
Clusters (Obs.)	1989 (5967)	1989 (5967)	1989 (5967)	1989 (5967)	1989 (5967)

Notes: * indicates significance at p<0.10, ** at p<0.05, *** at p<0.01, and **** at p<0.001. Demographic controls are a dummy for whether participant attended high school in the United States, dummies for each race category, dummies for each education category, fixed effects for round, average score in the category for the participant's own gender, and number of ASVAB questions answered correctly. Own gender advantage is the average gender difference in scores in the domain, signed so that a positive difference indicates an own gender advantage.

Table B4. Robustness of Posterior Beliefs to Other Approaches

	OLS Predicting Belief $Belief_{i,j,t}$					
	Posterior Beliefs					
	Belief of Score	**Mean of Full Posterior over Scores**	**Belief of Percentile Rank (1-100, 1 is Best)**	**Belief of Score**	**Mean of Full Posterior over Scores**	**Belief of Percentile Rank (1-100, 1 is Best)**
	I	II	III	IV	V	VI
$Female_i$	-0.33***	-0.37***	5.32***	-0.35***	-0.39***	5.47***
	(0.092)	(0.097)	(0.85)	(0.091)	(0.097)	(0.85)
Own Gender Advantage $(\overline{Score}_{j,G} - \overline{Score}_{j,-G})$	0.24***	0.24***	-1.65***			
	(0.022)	(0.021)	(0.20)			
Own Gender Advantage in Perception				1.02***	1.01***	-6.59***
				(0.091)	(0.092)	(0.87)
$Score_{i,j}$	0.79***	0.77***	-2.55***	0.79***	0.77***	-2.55***
	(0.012)	(0.012)	(0.11)	(0.012)	(0.012)	(0.11)
Demographic Controls	Yes	Yes	Yes	Yes	Yes	Yes
Adjusted R-squared	0.618	0.587	0.161	0.618	0.5871	0.1624
Clusters (Obs.)	1989 (5967)	1989 (5967)	1989 (5967)	1989 (5967)	1989 (5967)	1989 (5967)

Notes: * indicates significance at p<0.10, ** at p<0.05, *** at p<0.01, and **** at p<0.001. Demographic controls are a dummy for whether participant attended high school in the United States, dummies for each race category, dummies for each education category, fixed effects for round, average score in the category for the participant's own gender, and number of ASVAB questions answered correctly.

Table B5. Self-Stereotyping in Prior and Posterior Beliefs Using Score Fixed Effects

	OLS Predicting Belief of Score $Belief_{i,j,t}$			
	Prior Belief	Posterior Belief		
	Full Sample	Full Sample	50% Signal Accuracy	70% Signal Accuracy
	I	II	III	IV
$Female_i$	-0.54***	-0.42***	-0.28**	-0.54***
	(0.098)	(0.091)	(0.12)	(0.13)
Own Gender Advantage $(\overline{Score}_{j,G} - \overline{Score}_{j,-G})$	0.22***	0.16***	0.17***	0.15***
	(0.024)	(0.023)	(0.031)	(0.033)
$Score_{i,j}$ Fixed Effects	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes
Adjusted R-squared	0.4734	0.6423	0.6475	0.6457
Clusters (Obs.)	1989 (5967)	1989 (5967)	992 (2976)	997 (2991)

Notes: * indicates significance at p<0.10, ** at p<0.05, *** at p<0.01, and **** at p<0.001. Controls are a dummy for whether participant attended high school in the United States, dummies for each race category, dummies for each education category, fixed effects for round, average score in the category for the participant's own gender, and number of ASVAB questions answered correctly.

Table B6. Robustness of Systematic Deviations from the Bayesian Benchmark

	OLS Predicting Belief of Score $Belief_{i,j,t}$		
	Posterior Beliefs		
	Original	Smoothed Priors	Using Means
$Female_i$	-0.29***	-0.30***	-0.30***
	(0.083)	(0.083)	(0.086)
Own Gender Advantage $(\overline{Score}_{j,G} - \overline{Score}_{j,-G})$	0.15***	0.15***	0.15***
	(0.021)	(0.021)	(0.020)
$Score_{i,j}$	0.39***	0.38***	0.32***
	(0.017)	(0.017)	(0.021)
$Bayes_{i,j}$	0.45***		
	(0.016)		
Smoothed $Bayes_{i,j}$		0.47***	
		(0.016)	
$Bayes_{i,j}$ Computed with Means			0.52***
			(0.019)
Controls	Yes	Yes	Yes
Adjusted R-squared	0.6825	0.6861	0.667
Clusters (Obs.)	1989 (5967)	1989 (5967)	1989 (5967)

Notes: * indicates significance at p<0.10, ** at p<0.05, *** at p<0.01, and **** at p<0.001. Controls are a dummy for whether participant attended high school in the United States, dummies for each race category, dummies for each education category, fixed effects for round, average score in the category for the participant's own gender, and number of ASVAB questions answered correctly. In Column II, the Bayesian prediction is calculated using smoothed priors that place a small positive probability on every score that received zero probability in reported prior. In Column III, we use the mean of Bayesian-predicted posterior distribution as the Bayesian prediction, and we predict the mean of full posterior distribution rather than the posterior belief of score.

Table B7. Systematic Deviations from the Bayesian Benchmark using Score Fixed Effects

	OLS Predicting Belief of Score $Belief_{i,j,t}$	
	Posterior Beliefs	
	I	II
$Female_i$	-0.42***	-0.36***
	(0.091)	(0.084)
Own Gender Advantage $(\overline{Score}_{j,G} - \overline{Score}_{j,-G})$	0.16***	0.11***
	(0.023)	(0.022)
$Bayes_{i,j}$		0.41***
		(0.017)
$Score_{i,j}$ Fixed Effects	Yes	Yes
Controls	Yes	Yes
Adjusted R-squared	0.6423	0.6940
Clusters (Obs.)	1989 (5967)	1989 (5967)

Notes: * indicates significance at p<0.10, ** at p<0.05, *** at p<0.01, and **** at p<0.001. Controls are a dummy for whether participant attended high school in the United States, dummies for each race category, dummies for each education category, fixed effects for round, average score in the category for the participant's own gender, and number of ASVAB questions answered correctly.

Table B8. Responsiveness of Belief Updating by Gender and Gender Stereotypes

	OLS Predicting Belief of Score $Belief_{i,j,t}$			
	Posterior Beliefs			
	Male-Typed Domains		Female-Typed Domains	
	Men	Women	Men	Women
Smoothed $Bayes_{i,j}$ De-meaned	0.80***	0.61***	0.71***	0.83***
	(0.020)	(0.022)	(0.029)	(0.014)
Constant	7.09***	6.17***	6.81***	6.90***
	(0.080)	(0.076)	(0.089)	(0.060)
Adjusted R-squared	0.6408	0.468	0.533	0.734
Cluster (Obs.)	753 (1240)	1097 (1779)	716 (1139)	1123 (1836)

Notes: * indicates significance at p<0.10, ** at p<0.05, *** at p<0.01, and **** at p<0.001. Male-typed are categories that have an average positive value on slider scale rating (Cars, Sports, Videogames, Business); female-typed are categories that have an average negative value on slider scale rating (Kardashians, Disney, Art, Verbal). In these models, the Bayesian prediction is calculated using smoothed priors that place a small positive probability on every score that received zero probability in reported prior. The Bayesian predictions are then de-meaned by differencing out the mean Bayesian prediction in the full sample.

Table B9. Good News and Bad News Gender-Interacted Models

	OLS Predicting Belief of Score $Belief_{i,j,t}$			
	Posterior Beliefs			
	Male-Typed Domains		Female-Typed Domains	
	Good News	Bad News	Good News	Bad News
$Bayes_{i,j}$	0.79***	0.76***	0.68***	0.79***
	(0.022)	(0.052)	(0.034)	(0.055)
$Female_i$	-0.89***	-0.79***	0.043	0.45**
	(0.12)	(0.22)	(0.12)	(0.20)
$Bayes_{i,j}\ x\ Female_i$	-0.20***	-0.015	0.11***	0.13**
	(0.035)	(0.065)	(0.038)	(0.060)
Constant	7.04***	7.38***	6.81***	6.79***
	(0.084)	(0.18)	(0.10)	(0.15)
Clusters (Observations)	1645 (2412)	525 (580)	1640 (2396)	530 (579)
Adjusted R-squared	0.582	0.6399	0.656	0.7535

Notes: * indicates significance at p<0.10, ** at p<0.05, *** at p<0.01, and **** at p<0.001. Good news is a dummy that takes 1 if the signal received was greater than or equal to true score. Bayesian prediction is de-meaned.

Table B10. Good and Bad News Using Smoothed Priors

	OLS Predicting Belief of Score $Belief_{i,j,t}$			
	Posterior Beliefs			
	Male-Typed Domains		Female-Typed Domains	
	Men	Women	Men	Women
Smoothed $Bayes_{i,j}$	0.79***	0.76***	0.78***	0.93***
	(0.051)	(0.037)	(0.056)	(0.021)
$GoodNews_{i,j}$	-0.33*	-0.48***	0.061	-0.39***
	(0.19)	(0.15)	(0.18)	(0.13)
Smoothed $Bayes_{i,j}\ x\ GoodNews_{i,j}$	0.015	-0.17***	-0.085	-0.12***
	(0.056)	(0.045)	(0.065)	(0.026)
Constant	7.35***	6.62***	6.78***	7.24***
	(0.18)	(0.12)	(0.15)	(0.12)
Adjusted R-squared	0.6417	0.4731	0.5346	0.737
Cluster (Obs.)	735 (1213)	1097 (1779)	716 (1139)	1123 (1836)

Notes: * indicates significance at p<0.10, ** at p<0.05, *** at p<0.01, and **** at p<0.001. Good news is a dummy that takes 1 if the signal received was greater than or equal to true score. Male-typed are categories that have an average positive value on slider scale rating (Cars, Sports, Videogames, Business); female-typed are categories that have an average negative value on slider scale rating (Kardashians, Disney, Art, Verbal). In these models, the Bayesian prediction is calculated using smoothed priors that place a small positive probability on every score that received zero probability in reported prior. The Bayesian predictions are then de-meaned by differencing out the mean Bayesian prediction in the full sample.

Table B11. Good and Bad News Relative to Priors, Rather than True Score

	OLS Predicting Belief of Score $Belief_{i,j,t}$			
	Posterior Beliefs			
	Male-Typed Domains		Female-Typed Domains	
	Men	Women	Men	Women
$Bayes_{i,j}$	0.79***	0.81***	0.73***	0.90***
	(0.035)	(0.031)	(0.060)	(0.021)
$GoodNews_{i,j}$ rel. to Prior	-0.62***	-0.77***	-0.37**	-0.41***
	(0.18)	(0.14)	(0.18)	(0.12)
$Bayes_{i,j}$ x $GoodNews_{i,j}$ rel. to Prior	0.0038	-0.27***	-0.030	-0.11***
	(0.043)	(0.041)	(0.069)	(0.028)
Constant	7.55***	6.81***	7.05***	7.22***
	(0.16)	(0.11)	(0.15)	(0.11)
Adjusted R-squared	0.6385	0.4795	0.5247	0.7266
Cluster (Obs.)	735 (1213)	1097 (1779)	716 (1139)	1123 (1836)

Notes: * indicates significance at p<0.10, ** at p<0.05, *** at p<0.01, and **** at p<0.001. Good news is a dummy that takes 1 if the signal received was greater than or equal to true score. Male-typed are categories that have an average positive value on slider scale rating (Cars, Sports, Videogames, Business); female-typed are categories that have an average negative value on slider scale rating (Kardashians, Disney, Art, Verbal). The Bayesian predictions are de-meaned by differencing out the mean Bayesian prediction in the full sample.

Table B12. Correlates of Stereotype-Driven Updating

	OLS Predicting Belief of Score $Belief_{i,j,t}$			
	Posterior Beliefs			
	I	II	III	IV
$Female_i$	-0.33***	-0.29***	-0.036	-0.057
	(0.092)	(0.083)	(0.056)	(0.056)
Own Gender Advantage $(\overline{Score}_{j,G} - \overline{Score}_{j,-G})$	0.24***	0.15***	0.049***	0.019
	(0.022)	(0.021)	(0.016)	(0.016)
$Score_{i,j}$	0.79***	0.39***	0.39***	
	(0.012)	(0.017)	(0.015)	
$Bayes_{i,j}$		0.45***	0.027	0.00067
		(0.016)	(0.019)	(0.019)
Mode of Prior			0.62***	
			(0.020)	
Weight on Mode in Prior			0.0028**	0.0015
			(0.0013)	(0.0013)
Weight on Signal in Prior			0.00021	-0.00012
			(0.00099)	(0.00097)
Std. Dev. Of Prior			0.099***	0.087***
			(0.032)	(0.033)
Right-Skewedness			0.27***	0.26***
			(0.066)	(0.066)
Noise			0.24***	0.25***
			(0.015)	(0.015)
Constant	2.86***	2.04***	-0.28	0.21
	(0.73)	(0.71)	(0.58)	(0.61)
Score Fixed Effects	No	No	No	Yes
Prior Mode Fixed Effects	No	No	No	Yes
Controls	Yes	Yes	Yes	Yes
Clusters (Observations)	1989 (5967)	1989 (5967)	1989 (5967)	1989 (5967)
Adjusted R-squared	0.618	0.681	0.815	0.820

Notes: * indicates significance at p<0.10, ** at p<0.05, *** at p<0.01, and **** at p<0.001. Noise draw is the random draw of an integer drawn from [-5,5] that was added to the score to determine the signal observed. Controls are a dummy for whether participant attended high school in the United States, dummies for each race category, dummies for each education category, fixed effects for round, average score in the category for the participant's own gender, and number of ASVAB questions answered correctly.

Table B13. Self-Stereotyping in Prior and Posterior Beliefs for Restricted Domains

	OLS Predicting Belief of Score $Belief_{i,j,t}$		
	Using Only Business, Verbal, and Art		
	Prior Beliefs	Posterior Beliefs	
	I	II	III
$Female_i$	-0.28**	-0.13	-0.076
	(0.14)	(0.13)	(0.12)
Own Gender Advantage $\left(\overline{Score}_{j,G} - \overline{Score}_{j,-G}\right)$	0.23**	0.28***	0.21**
	(0.11)	(0.10)	(0.097)
$Score_{i,j}$	0.53***	0.69***	0.29***
	(0.024)	(0.023)	(0.027)
$Bayes_{i,j}$			0.49***
			(0.029)
Controls	Yes	Yes	Yes
Adjusted R-squared	0.2943	0.4573	0.5617
Clusters (Obs.)	1631 (2230)	1631 (2230)	1631 (2230)

Notes: * indicates significance at p<0.10, ** at p<0.05, *** at p<0.01, and **** at p<0.001. Controls are a dummy for whether participant attended high school in the United States, dummies for each race category, dummies for each education category, fixed effects for round, average score in the category for the participant's own gender, and number of ASVAB questions answered correctly. Sample is restricted to only observations from Business, Verbal Skills, and Art and Literature.

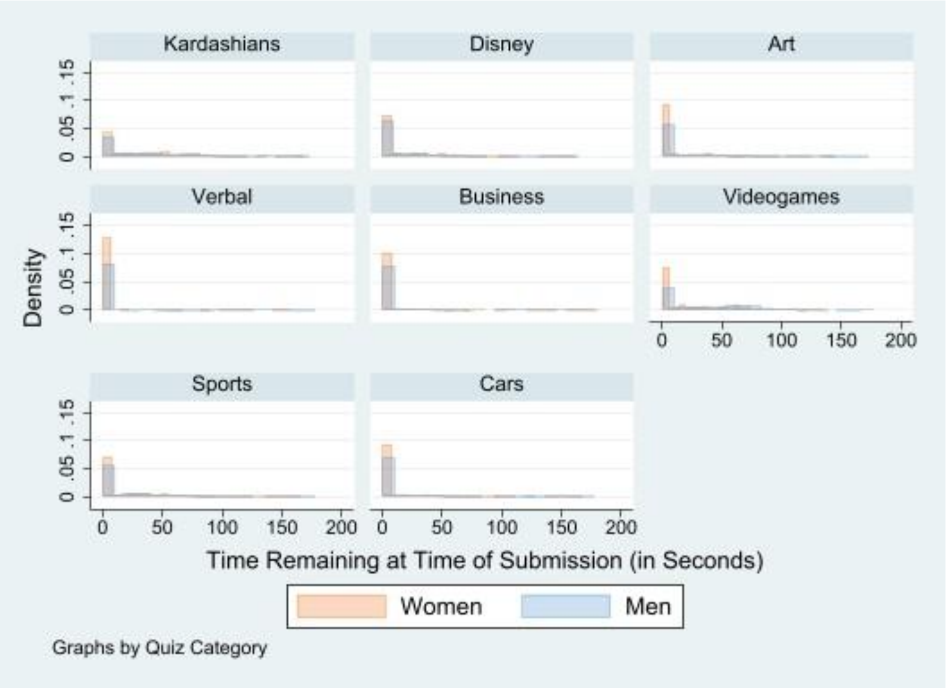


Figure B1. Distribution of Time Remaining at Quiz Submission

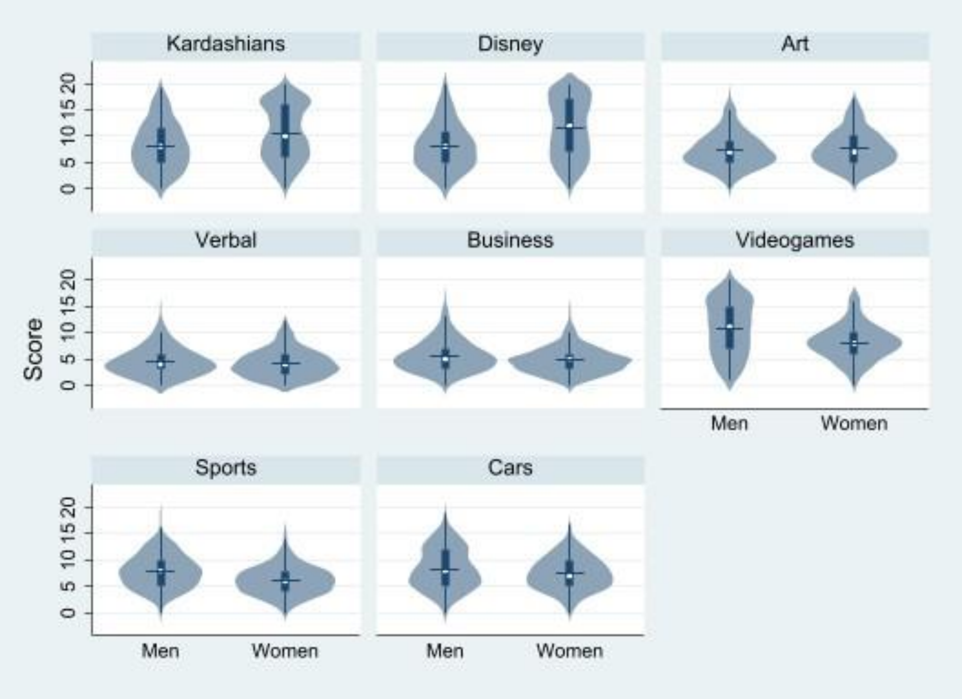


Figure B2. Distribution of Scores by Gender and Category

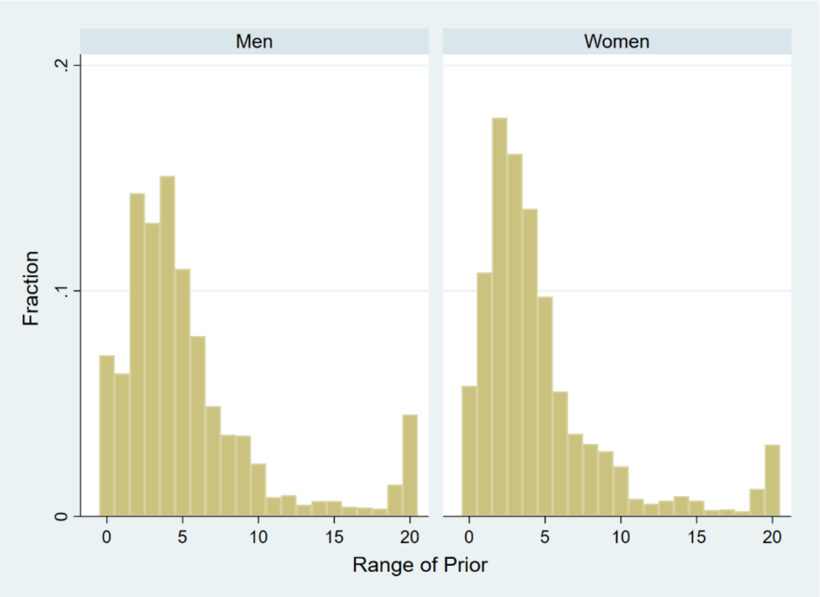


Figure B3. Range of Priors by Gender

APPENDIX C.

While the experiment was running on Amazon Mechanical Turk, we noticed an error in the instructions for why participants had an incentive to tell the truth for the distribution elicitation questions. Note that corrected instructions are available in Appendix F. The error related to the description of "Payment Option 2: The Bet On Your Score". Essentially, instead of offering participants a payment if their score was equal to the score they guessed (the correct structure), the instructions incorrectly copied the "Payment Option 1: The Lottery" language, suggesting that a random integer would be compared to their percentage. See the text below, including the highlight in yellow for the error.

Payment Option 2: The Bet on Your Score

In the question above, you will tell us that you think there is a ____ % chance of your true score being equal to a given value. Let's call this value your "Percentage". That is, if you tell us that you think there is a 60% chance your true score is equal to the value, then your percentage is 60.

Then, we will draw a second an integer at random between 1- 100. Again, each integer (1,2,3,4,..., 96, 97, 98, 99, 100) is equally likely to be chosen. We'll call this number that's chosen the "Draw".

If the "Draw" number is less than or equal to your "Percentage" number, the lottery will pay you $1. If not, that is, if the "Draw" number is more than the "Percentage" number, the lottery will pay you nothing.

Under this procedure, there is no clear financial incentive for truth-telling. The first 1647 participants saw this error in the instructions, while the final 374 participants viewed corrected instructions. We can test to see whether this error appears to impact the answers participants gave. The cleanest test is comparing the answers across participants for whom the error was fixed or not fixed. We can do this in a few ways. In Column 1 of Table C1 below, we can compare first responses across participants – that is, look at answers for the first question in which they saw these instructions (eliciting the probability assigned to the particular mode of their prior in the first round). First responses may be most reasonable because participants had to choose to view these instructions, and it is likely rates of viewing these instructions decrease with each opportunity to view them. In Column 2, we compare all responses to questions about the probability mass assigned to the mode of their prior. In Column 3, we compare all responses to questions about the probability mass assigned to the mode of their posterior. In each specification, the dummy for the error in the instructions is insignificant and the point estimate is close to 0 (recall that the outcome variable ranges from 0 – 100).

Table C1. Evaluating the Error in the Instructions

	OLS Predicting Probability Assigned to Mode of Prior – First Observation	**OLS Predicting Probability Assigned to Mode of Prior – All Observations**	**OLS Predicting Probability Assigned to Mode of Posterior**
Instruction Error was Fixed	-1.83 (1.334)	-1.32 (1.050)	0.07 (1.023)
Demographic Controls	Yes	Yes	Yes
R-squared	0.02	0.02	0.01
Cluster (Obs.)	1989 (1989)	1989 (5967)	1989 (5967)

Notes: * indicates significance at p<0.10, ** at p<0.05, *** at p<0.01, and **** at p<0.001. Demographic controls are a dummy for whether participant attended high school in the United States, dummies for each race category, dummies for each education category, fixed effects for round (columns 3 and 4), average score in the category for the participant's own gender, and number of ASVAB questions answered correctly.

APPENDIX D. THEORETICAL FRAMEWORK

Denote the decision-maker's true score on the test by T, where $T \in \{0,1,2, \ldots, 30\}$. The decision-maker holds a prior belief over the distribution of possible test scores, such that for each possible test score, t, the decision-maker believes she earned that test score with probability, r(T=t), where:

$$\sum_{t=0}^{t=30} \mathrm{r}(\mathrm{T}=\mathrm{t})=1$$

The mode of that prior distribution is t' such that $r(T = t') > r(T = t)$ for all $t \neq t'$.[19] The decision-maker then receives a signal of her performance, X, where $X \in \{T-5, T-4, T-3, T-2, T-1, T, T+1, T+2, T+3, T+4, T+5\}$. We have that X = T with probability q, where q varies with treatment assignment.[20] After viewing the signal, the decision-maker then forms a posterior belief over the distribution of possible test scores, such that for each possible test score, t, the decision-maker believes she earned that test score with probability, s(T=t), where:

$$\sum_{t=0}^{t=30} \mathrm{s}(\mathrm{T}=\mathrm{t})=1$$

The mode of that posterior distribution is t* such that $s(T = t^*) > s(T = t)$ for all $t \neq t^*$.[21]

What can we say about the beliefs a Bayesian decision-maker would hold after observing a signal in our framework? Conditional on having a score of T, she observes signal X = T with probability q. And, for each other score, $t \in \{T-5, T-4, T-3, T-2, T-1, T, T+1, T+2, T+3, T+4, T+5\}$, she observes signal X = t with probability (1-q)/10.[22] Suppose now that the decision-maker observes a signal X = Z. This signal can be generated by 11 possible true scores. A true score of T=Z generates this signal with probability q. Any other true score in {Z-5, Z-4, Z-3, Z-2, Z-1, Z+1, Z+2, Z+3, Z+4, Z+5} generates this signal with probability (1-q)/10. No other true score can generate the observed signal X = Z. This implies that signal X = Z has been

[19] The case where no such t' exists, due to the decision-maker assigning equal likelihood to the two (or more) most likely scores, occurs with probability 0.

[20] In the motivating study (Appendix A), we explore q = 0.6 and q=0.9. In the main study, we explore q = 0.5 and q = 0.7.

[21] Again, the case where no such t* exists, due to the decision-maker assigning equal likelihood to the two (or more) most likely scores, occurs with probability 0.

[22] For each of these other scores, the signal process generates an incorrect signal with probability (1-q), and each incorrect score in the feasible range occurs with equal probability.

generated by a score of T = Z with probability q and by T = Z+i with probability (1-q)/10 for each i ϵ {-5,-4,-3,-2,-1,1,2,3,4,5}.

Now consider the role of her prior. We can use Bayes rule to write down an expression for a decision-maker's posterior probability of holding any particular score, given her prior belief distribution and the signal she has received. Denote the probability of observing the signal X = Z conditional on T = t by:

p(X = Z|T=t)

First, let's consider her posterior belief, s(T = t) for the case where t = Z. That is, having seen a particular signal, what will be the decision-maker's posterior belief of her true score being equal to that signal?

$$s(T = t = Z) = \frac{\mathrm{p(X\ =\ Z|T = t = Z) \times\ r(T = t = Z)}}{\mathrm{(\ p(Z = t|T = t = Z)\ \times\ r(T = t = Z)\ )}\ + \sum_{i=-5}^{i=5} \mathrm{p(X\ =\ Z|T = Z + i)\ \times\ r(T\ =\ Z\ +\ i)}}$$

We can use the probabilities computed above, in particular p(X=Z|T=t=Z) = q and p(X=Z|T=Z+i) = (1-q)/10 for each i ϵ {-5,-4,-3,-2,-1,1,2,3,4,5}, to produce:

$$s(T = t = Z) = \frac{\mathrm{q\ r(T = t = Z)}}{\mathrm{q\ r(T = t = Z)}\ + \sum_{i=-5}^{i=5} \frac{(1-q)}{10} \times \mathrm{r(T\ =\ Z\ +\ i)}}$$

Of course, the same formula can be used to produce her posterior belief of holding any particular score, t≠Z, after having seen signal X=Z. In cases where Z ≠ t, we have:

$$s(T = t \neq Z) = \frac{\mathrm{p(X\ =\ Z|T = t \neq Z) \times\ r(T = t \neq Z)}}{\mathrm{(\ p(Z = t|T = t = Z)\ \times\ r(T = t = Z)\ )}\ + \sum_{i=-5}^{i=5} \mathrm{p(X\ =\ Z|T = Z + i)\ \times\ r(T\ =\ Z\ +\ i)}}$$

First note that for all t such that t does not fall within {Z-5, Z-4, Z-3, Z-2, Z-1, Z, Z+1, Z+2, Z+3, Z+4, Z+5}, we have p(X = Z|T=t) = 0, and thus s(T = t) = 0. In words, a Bayesian cannot justify placing positive probability on a score that could not have generated the observed signal X = Z. For all t in {Z-5, Z-4, Z-3, Z-2, Z-1, Z, Z+1, Z+2, Z+3, Z+4, Z+5}, we can sub in using the probabilities above to get:

$$s(T = t = Z) = \frac{\frac{(1-q)}{10}\ \mathrm{r(T = t \neq Z)}}{\mathrm{q\ r(T = Z)}\ + \sum_{i=-5}^{i=5} \frac{(1-q)}{10} \times \mathrm{r(T\ =\ Z\ +\ i)}}$$

With these formulas, we can determine what the mode of a Bayesian's posterior distribution should be, as a function of the signal observed and her prior beliefs. The first natural question to ask is, when will the mode of a Bayesian's posterior be the signal she observed? That is, given X = Z, when will t* = Z? In order for this *not* to be the case, we would need:

$$\exists t \in \{\mathrm{Z-5, Z-4, Z-3, Z-2, Z-1, Z, Z+1, Z+2, Z+3, Z+4, Z+5}\}$$
$$such\ that$$
$$\mathrm{s(T\ =\ Z)\ <\ s(T\ =\ t)}$$

Plugging in,

$$\frac{\mathrm{q\ r(T = t = Z)}}{\mathrm{q\ r(T = t = Z)}\ + \sum_{i=-5}^{i=5} \frac{(1-q)}{10} \times \mathrm{r(T\ =\ Z\ +\ i)}} < \frac{\frac{(1-q)}{10}\ \mathrm{r(T = t \neq Z)}}{\mathrm{q\ r(T = Z)}\ + \sum_{i=-5}^{i=5} \frac{(1-q)}{10} \times \mathrm{r(T\ =\ Z\ +\ i)}}$$

Simplifying,

$$qr(T = Z) < \frac{(1-q)}{10} r(T = t)$$
$$r(T = t) > \frac{10q\ r(T = Z)}{(1-q)}$$

This tells us that, in order for the signal received, Z, to *not* be the mode of the posterior, it must be the case that the decision-maker placed sufficiently little probability on her true score being equal to Z in her prior, relative to the probability she placed on at least one other score (that is feasible given the signal received). We can use the probabilities, q={0.5, 0.6, 0.7, 0.9}, from our experiments to reach the following propositions.

Proposition 1. Suppose a decision-maker observes X = Z in the 50% signal accuracy treatment. Then, unless exists t such that r(T=t) > 10r(T=Z) with p(X = Z|T=t) > 0, it must be the case that the mode of her posterior is the signal she observed; that is, t* = Z. Suppose a decision-maker observes X = Z in the 60% signal accuracy treatment. Then, unless exists t such that r(T=t) > 15r(T=Z) with p(X = Z|T=t) > 0, it must be the case that the mode of her posterior is the signal she observed; that is, t* = Z. Suppose a decision-maker observes X = Z in the 70% signal accuracy treatment. Then, unless exists t such that r(T=t) > (70/3)r(T=Z) with p(X = Z|T=t) > 0, it must be the case that the mode of her posterior is the signal she observed; that is, t* = Z. Suppose a decision-maker observes X = Z in the 90% signal accuracy treatment. Then, unless exists t such that r(T=t) > 90r(T=Z) with p(X = Z|T=t) > 0, it must be the case that the mode of her posterior is the signal she observed; that is, t* = Z. Given that prior probabilities must sum to 1, this also implies that if r(T=Z)> 1/11 in the 50% treatment, r(T=Z)> 1/16 in the 60% treatment, r(T=Z)> 3/73 in the 70% treatment, or r(T=Z)>1/91 in the 90% treatment, then it must be that t* = Z.

Because of the informativeness of our signals, the mode of a Bayesian's posterior will be her signal, except in cases where she put very little weight on the signal being her true score in her prior. In those cases, what will be the mode of her posterior? We show below that, if the mode of the posterior is not the signal received, Z, then it must be the case that the mode of the posterior, t*, is the mode of the prior over {Z-5, Z-4, Z-3, Z-2, Z-1, Z, Z+1, Z+2, Z+3, Z+4, Z+5}. There are two cases to consider. In the first case, t' in {Z-5, Z-4, Z-3, Z-2, Z-1, Z, Z+1, Z+2, Z+3, Z+4, Z+5}. That is, the mode of the decision-maker's prior could have feasibly generated the signal observed.

Then, in this case, for it to be true that the mode of the decision-maker's prior is *not* the mode of the decision-maker's posterior, there would have to exist some t_j, where $t_j \neq Z$ and $t_j \neq t'$, such that s(T = t_j) > s(T = t'). Because we know $t_j \neq Z$, this implies:

$$\frac{\frac{(1-q)}{10}\,\mathrm{r}(\mathrm{T}=t_j)}{\mathrm{q}\,\mathrm{r}(\mathrm{T}=\mathrm{Z})\;+\sum_{i=-5}^{i=5}\frac{(1-q)}{10}\times\mathrm{r}(\mathrm{T}\;=\;\mathrm{Z}\;+\;\mathrm{i})} > \frac{\frac{(1-q)}{10}\,\mathrm{r}(\mathrm{T}=\mathrm{t}')}{\mathrm{q}\,\mathrm{r}(\mathrm{T}=\mathrm{Z})\;+\sum_{i=-5}^{i=5}\frac{(1-q)}{10}\times\mathrm{r}(\mathrm{T}\;=\;\mathrm{Z}\;+\;\mathrm{i})}$$

Or, more simply:

$$\mathrm{r}\big(\mathrm{T}=t_j\big) > \mathrm{r}(\mathrm{T}=\mathrm{t}')$$

But, this is a contradiction, as t' is the mode of the prior. Thus, it must be that if t* $\neq$ Z and t' in {Z-5, Z-4, Z-3, Z-2, Z-1, Z, Z+1, Z+2, Z+3, Z+4, Z+5}, it must be that t* = t'.

This leaves one remaining case, the case in which exists t in {Z-5, Z-4, Z-3, Z-2, Z-1, Z, Z+1, Z+2, Z+3, Z+4, Z+5} such that r(T=t) > ((1-q)/10)r(T=Z), so that the mode of the posterior is not the signal received, *and,* the mode of the decision-maker's prior could not have generated her observed signal, t'< Z-5 or t'>

Z+5. In these cases, the decision-maker should report as the mode of her posterior the value t_j such that: s(T =t_j) > s(T = t_k) for all k $\neq$ j and t_j, t_k in {Z-5, Z-4, Z-3, Z-2, Z-1, Z, Z+1, Z+2, Z+3, Z+4, Z+5}.

Plugging in,

$$\frac{\frac{(1-q)}{10}\,\mathrm{r}(\mathrm{T}=t_j)}{\mathrm{q}\,\mathrm{r}(\mathrm{T}=\mathrm{Z})\;+\sum_{i=-5}^{i=5}\frac{(1-q)}{10}\times\,\mathrm{r}(\mathrm{T}\;=\;\mathrm{Z}\;+\;\mathrm{i})} > \frac{\frac{(1-q)}{10}\,\mathrm{r}(\mathrm{T}=t_k)}{\mathrm{q}\,\mathrm{r}(\mathrm{T}=\mathrm{Z})\;+\sum_{i=-5}^{i=5}\frac{(1-q)}{10}\times\,\mathrm{r}(\mathrm{T}\;=\;\mathrm{Z}\;+\;\mathrm{i})}$$

Or, more simply:

$$\mathrm{r}\big(\mathrm{T}=t_j\big) > \mathrm{r}(\mathrm{T}=t_k)$$

In this case, the decision-maker should report the t in {Z-5, Z-4, Z-3, Z-2, Z-1, Z, Z+1, Z+2, Z+3, Z+4, Z+5} for which r(T=t) is largest. Or, put differently, the decision-maker should report as the mode of her posterior the mode of her prior restricted to the distribution over {Z-5, Z-4, Z-3, Z-2, Z-1, Z, Z+1, Z+2, Z+3, Z+4, Z+5}. This leads to the following proposition that fully covers the two cases in which t* $\neq$ Z.

Proposition 2. Suppose t* $\neq$ Z. Then, the mode of the posterior is the mode of the prior distribution restricted to {Z-5, Z-4, Z-3, Z-2, Z-1, Z, Z+1, Z+2, Z+3, Z+4, Z+5}. In the event that t' in {Z-5, Z-4, Z-3, Z-2, Z-1, Z, Z+1, Z+2, Z+3, Z+4, Z+5}, then t* = t'.

APPENDIX E. ALTERNATIVE FRAMEWORK

Benjamin (2019) organizes his review of the belief updating literature around a model introduced by Grether (1980). In this section, we re-visit our analysis through the lens of this alternative framework. For simplicity, and to increase the extent to which our results are easily compared with results from other settings, we focus our analysis around two (of many) possible states of the world, *A* and *B.* Let *A* be the state of the world in which an individual's true score is equal to the mode of her prior belief. Let *B* be the state of the world in which an individual's true score is equal to the signal she receives. In our framework, A and B are not necessarily mutually exclusive, as an individual could receive a signal equal to the mode of her prior belief.

The decision maker's problem is to determine the relative likelihood of states A and B after observing a signal, *Z*. Using the Grether (1980) model, we can write:

$$\frac{s(A|Z)}{s(B|Z)} = \left[\frac{p(Z|A)}{p(Z|B)}\right]^c \left[\frac{r(A)}{r(B)}\right]^d$$

where *s* is the posterior belief of the state, conditional on the signal received, *r* is the individual's prior belief of the likelihood of the state, $p(Z|A)$ is the likelihood of receiving signal *Z* conditional on state *A*, and $p(Z|B)$ is the likelihood of receiving signal *Z* conditional on state *B*. The parameters *c* and *d* describe deviations from the Bayesian benchmark related to biased use of likelihoods and biased use of priors, respectively.

Taking logs, we have:

$$\ln\frac{s(A|Z)}{s(B|Z)} = c\ln\left[\frac{p(Z|A)}{p(Z|B)}\right] + d\ln\left[\frac{r(A)}{r(B)}\right]$$

Note that there are some challenges in producing these variables with our data. In particular, we have a non-trivial amount of extreme subjective beliefs: believed probabilities of 0 or 1. To deal with this, we winsorize our data. In particular, for those participants who report a subjective belief (for s or r) less than 0.01, we replace these values with 0.01. And, for those participants who report a subjective belief (for s or r) equal to 1, we replace these values with 0.99.[23]

Computing ln $\left[\frac{p(Z|A)}{p(Z|B)}\right]$ is somewhat easier, as these probabilities are objectively determined by our paradigm. However, a significant fraction of our participants (21%) provide a mode of their prior that could not generate the observed signal. For those participants, $p(Z|A)$=0. A Bayesian should assign no weight to this prior mode in their posterior, as there is zero probability that this score generated the observed signal. In order to provide results for our full sample, we replace $p(Z|A)$=0 with $p(Z|A)$=0.01 for those participants in our full sample specifications.

Gender Congruence and Deviations from the Bayesian Benchmark

Table E.1 presents the estimates of c and d from our data. Consistent with past work, we find significant under-inference (*c<1)* and significant base-rate neglect (*d<1*). Using the winsorized data, we estimate that biased use of likelihoods plays a relatively larger role than biased use of priors in describing deviations from Bayesian predictions (*c<d).* This is true independent of whether we use the winsorized full sample (Column I), or the winsorized sample that excludes individuals whose prior mode could not have generated the signal they observed (Column II). When we do not winsorize the data, restricting attention to only those observations for whom all three variables are well-defined without winsorizing, we estimate more severe under-inference and base-rate neglect, with a reversing of their relative roles (Column III). However, we would urge caution in interpreting this specification given that it uses only approximately one third of our data.

[23] Specifically, we re-code the 25% of observations for whom $s(A|Z)$<0.01 as 0.01, and the 5% of observations for whom $s(A|Z) = 1$ as 0.99. We-recode the 35% of observations for whom $s(B|Z)$<0.01 as 0.01, and the 8% of observations for whom $s(B|Z) = 1$ as 0.99. In terms of prior beliefs, we re-code the 0.17% of observations for whom $r(A)$<0.01 as 0.01, and the 4% of observations for whom $r(A) = 1$ as 0.99. We-recode the 53% of observations for whom $r(B)$<0.01 as 0.01, and the 1% of observations for whom $r(B) = 1$ as 0.99.

Table E.1 Identifying Deviations from the Bayesian Benchmark

	OLS Predicting Log of Posterior Beliefs $\ln \frac{s(A\|Z)}{s(B\|Z)}$			
		I	II	III
		Winsorized Full Sample	Winsorized Sample Restricted to $p(Z\|A)$>0	Non-Winsorized Sample
	Parameter			
$\ln\left[\frac{p(Z\|A)}{p(Z\|B)}\right]$	*c*	0.37**** (0.031)	0.58**** (0.032)	0.28**** (0.025)
$\ln\left[\frac{r(A)}{r(B)}\right]$	*d*	0.62**** (0.027)	0.62**** (0.028)	0.17**** (0.047)
Constant		-0.57**** (0.053)	-0.21**** (0.045)	-0.022 (0.019)
R-squared		0.096	0.11	0.044
Clusters (N)		1,989 (5,967)	1,962 (4,742)	1,334 (2,103)

Notes: Standard errors clustered at the individual level. Column I winsorizes observations less than 0.01 or greater than 0.99. Column II uses the same winsorization, but excludes observations for which $p(Z|A)$=0. Column III does not winsorize the data, excluding any observation for which a variable cannot be defined.
* indicates significance at p<0.10, ** at p<0.05, *** at p<0.01, and **** at p<0.001.

Of course, the main focus of our analysis is whether there is a role for the gender congruence of the domain in shaping belief updating. We can re-visit this key question using this alternative framework. We adapt the approach of Mobius et al (2022), who ask whether the degree of under-inference (*c)* depends upon whether the signal was good or bad news. Here, we ask whether the degree of under-inference depends upon whether the domain was gender congruent or gender incongruent. To do so, we construct the interaction of $\ln\left[\frac{p(Z|A)}{p(Z|B)}\right]$ with a dummy variable indicating whether the domain the observation is drawn from one was gender congruent for the participant, and, separately, the interaction of $\ln\left[\frac{p(Z|A)}{p(Z|B)}\right]$ with a dummy variable indicating whether the domain the observation is drawn from one was gender *incongruent* for the participant. Estimating a model that includes these two interaction terms and the log odds of prior beliefs, without a constant, can let us investigate whether the degree of under-inference depends upon the gender congruence of the domain.

Table E.2 presents the results. Consistent with our results from our main analysis, we estimate significantly greater average deviations from the Bayesian model in gender incongruent domains, as compared to gender congruent domains. This is true whether we use the winsorized full sample, or the restricted winsorized sample (Columns I and II). However, this result is only directionally true among the highly restricted sample in Column III. Overall, the results suggest that the extent of under-inference is significantly larger when signals arrive in incongruent domains.

Table E.2 Gender Congruence and the Extent of Under-Inference

	OLS Predicting Log of Posterior Beliefs $\ln\frac{s(A\mid Z)}{s(B\mid Z)}$			
		I	II	III
		Winsorized Full Sample	Winsorized Sample Restricted to $p(Z\mid A)$>0	Non-Winsorized Sample
	Parameter			
$Congruent\ x\ \ln\left[\frac{p(Z\mid A)}{p(Z\mid B)}\right]$	c_1	0.57**** (0.028)	0.70**** (0.031)	0.29**** (0.032)
$Incongruent\ x\ \ln\left[\frac{p(Z\mid A)}{p(Z\mid B)}\right]$	c_2	0.47**** (0.030)	0.60**** (0.032)	0.27**** (0.031)
$\ln\left[\frac{r(A)}{r(B)}\right]$	d	0.59**** (0.027)	0.61**** (0.028)	0.16**** (0.046)
p-value for F-test of $c_1 = c_2$		<0.001	0.001	0.59
R-squared		0.094	0.11	0.044
Clusters (N)		1,989 (5,967)	1,962 (4,742)	1,334 (2,103)

Notes: Standard errors clustered at the individual level. Column I winsorizes observations less than 0.01 or greater than 0.99. Column II uses the same winsorization, but excludes observations for which $p(Z|A)$=0. Column III does not winsorize the data, excluding any observation for which a variable cannot be defined.
* indicates significance at p<0.10, ** at p<0.05, *** at p<0.01, and **** at p<0.001.

Our primary analysis revealed that responses to good and bad news depended on the gender congruence of the domain. We can also explore this finding in the context of this alternative framework. In particular, we can expand the model of Table E.2 to further consider the interactions of gender congruence (and incongruence) with good and bad news. We define good and bad news as we did in the main text, where a participant is said to have received good news if her signal met or exceeded her true score, and bad news otherwise.

Table E.3 presents the results. We find substantial support for our finding from our main analysis. That is, we estimate that there is significantly more responsiveness to good news when it arrives in a gender congruent domain than when it arrives in a gender incongruent domain. In both Columns I and II, we estimate that the extent of under-inference (c) after seeing good news is significantly larger in incongruent domains than congruent domains. This is directionally true but not significantly so in the highly restricted sample of Column III.

Columns I and II also offer some support for asymmetry in reactions to good and bad news. We estimate that there is significantly greater responsiveness to good news than bad within gender congruent domains. This asymmetry is weaker within gender incongruent domains, where reactions to good news are more muted.

Table E.3 Gender Congruence, Good and Bad News, and the Extent of Under-Inference

	OLS Predicting Log of Posterior Beliefs $\ln\frac{s(A\mid Z)}{s(B\mid Z)}$			
		I	II	III
		Winsorized Full Sample	Winsorized Sample Restricted to $p(Z\mid A)$>0	Non-Winsorized Sample
	Parameter			
$Congruent\ x\ \ln\left[\frac{p(Z\mid A)}{p(Z\mid B)}\right]\ x\ Good\ News$	c_{1G}	0.59**** (0.030)	0.72**** (0.033)	0.29**** (0.033)
$Congruent\ x\ \ln\left[\frac{p(Z\mid A)}{p(Z\mid B)}\right]\ x\ Bad\ News$	c_{1B}	0.46**** (0.045)	0.60**** (0.053)	0.29**** (0.054)
$Incongruent\ x\ \ln\left[\frac{p(Z\mid A)}{p(Z\mid B)}\right]\ x\ Good\ News$	c_{2G}	0.47**** (0.031)	0.61**** (0.034)	0.26**** (0.032)
$Incongruent\ x\ \ln\left[\frac{p(Z\mid A)}{p(Z\mid B)}\right]\ x\ Bad\ News$	c_{2B}	0.44**** (0.046)	0.53**** (0.050)	0.32**** (0.054)
$\ln\left[\frac{r(A)}{r(B)}\right]$	d	0.58**** (0.027)	0.61**** (0.028)	0.16**** (0.046)
p-value for F-test of $c_{1G} = c_{2G}$		<0.001	0.003	0.44
p-value for F-test of $c_{1B} = c_{2B}$		0.61	0.28	0.72
p-value for F-test of $c_{1G} = c_{1B}$		0.005	0.03	0.92
p-value for F-test of $c_{2G} = c_{2B}$		0.40	0.10	0.29
R-squared		0.096	0.12	0.10
Clusters (N)		1,989 (5,967)	1,962 (4,742)	1,334 (2,103)

Notes: Standard errors clustered at the individual level. Column I winsorizes observations less than 0.01 or greater than 0.99. Column II uses the same winsorization, but excludes observations for which $p(Z|A)$=0. Column III does not winsorize the data, excluding any observation for which a variable cannot be defined.
* indicates significance at p<0.10, ** at p<0.05, *** at p<0.01, and **** at p<0.001.

Overall, this analysis offers further support for our main conclusions. In particular, we estimate greater deviations from the Bayesian model in gender incongruent domains. This seems driven largely by the fact that individuals are less responsive to good news when it arrives in an incongruent domain than when it arrives in a congruent domain.

Confirmation Bias (or Prior-Based Inference)

Past work has found evidence for confirmation bias, or, as Benjamin (2019) describes it, prior-based inference. This is the idea that participants infer more from confirmatory signals – signals in line with their prior beliefs. One plausible explanation for our results is that individuals under-infer less in gender congruent domains because they are more likely to receive confirmatory signals in gender congruent domains. We explore that hypothesis here.

Benjamin (2019), following Charness and Dave (2017), define a confirming signal as one such that $\frac{r(A)}{r(B)} > 1$ and $\frac{p(Z|A)}{p(Z|B)} > 1$, or $\frac{r(A)}{r(B)} < 1$ and $\frac{p(Z|A)}{p(Z|B)} < 1$, and a disconfirming signal as one such that $\frac{r(A)}{r(B)} > 1$ and $\frac{p(Z|A)}{p(Z|B)} < 1$, or $\frac{r(A)}{r(B)} < 1$ and $\frac{p(Z|A)}{p(Z|B)} > 1$. Under this classification, it is clear that nearly all of our signals are disconfirming.

First, $\frac{r(A)}{r(B)} \geq 1$. That is, the prior likelihood assigned to the mode of the prior distribution must equal or exceed the prior likelihood assigned to the signal received.[24] Second, given the signal structure, in all cases, we have $\frac{p(Z|A)}{p(Z|B)} \leq 1$. Unless the individual's prior mode is equal to the signal received, we have $p(Z|A) < p(Z|B)$.

Thus, for the vast majority of participants, we have that $\frac{r(A)}{r(B)} > 1$ and $\frac{p(Z|A)}{p(Z|B)} < 1$; for nearly all of the rest, we have that the signal is equal to the mode of the prior beliefs distribution. In fact, 86%

[24] In fact, in our data, this is not true for 76 observations. In those 76 cases, the score that was guessed as the most likely score in the first part of the elicitation does not in fact receive as much weight as the signal in the prior belief distribution. We do not force participants to be consistent across parts of the belief elicitation, so in practice they could provide one score as their most likely score, and yet not assign the most weight to that score when providing their full distribution. This is rare, occurring in less than 0.02% of cases (76/5,967).

of our signals are disconfirming. The rate of disconfirming signals does not differ by whether or not the domain is gender congruent (86.5% for congruent, 86.7% for incongruent, p=0.83). Thus, our results cannot be explained by individuals being more likely to receive confirmatory signals in gender congruent domains.

Of course, one could push even farther and ask whether the extent to which signals are disconfirming varies by the gender congruence of the domain. That is, does the difference in, or the ratio of, $\frac{r(A)}{r(B)}$ and $\frac{p(Z|A)}{p(Z|B)}$ depend upon the gender congruence of the domain? This is also not the case. If we construct a continuous measure of the extent to which a signal is disconfirming, we continue to see no differences in this measure across congruent and incongruent domains.

APPENDIX F. Experimental Materials.

F.1. Experimental Instructions Cognitive Skills Study (under separate cover)

https://perma.cc/55RY-UW87

F.2. Experimental Instructions Main Study (under separate cover)

https://perma.cc/5SC2-8MXH