HARVARD | BUSINESS | SCHOOL

Stereotypes and Belief Updating

Katherine Coffman
Manuela Collis
Leena Kulkarni

Working Paper 19-068

Stereotypes and Belief Updating

Katherine Coffman
Harvard Business School

Manuela Collis
Harvard Business School

Leena Kulkarni
Harvard School of Public Health

Working Paper 19-068

STEREOTYPES AND BELIEF UPDATING*

Katherine Coffman[‡]

Manuela Collis

Leena Kulkarni

June 2021

Abstract: We explore how feedback shapes, and perpetuates, gender gaps in self-assessments. Participants in our experiments 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. We thank the NSF 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.

[‡] 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 likely 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 each of these 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 2019 on self-presentation; 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 2018), 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).

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.

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 ability. We ask whether and how biased beliefs persist in the face

[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).

of highly informative, task-specific feedback. In particular, we run a series of experiments that explore how individuals update their beliefs about themselves in response to feedback about own ability. 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?

In our first experiment, participants complete a timed test of cognitive ability. After completing the test, we elicit an incentivized belief of their absolute score on the test from all participants. Then, we provide information: a noisy signal of their true score. Across two randomly-assigned conditions, we vary the precision of the signal received: half of our participants receive a signal that is equal to their true score with probability 0.6 and half receive a signal that is equal to their true score with probability 0.9. In both treatments, signals are equal to a participant's true score with probability *p* (either 0.6 or 0.9 depending on 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. Note that this 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. After transmitting these signals, we again elicit an incentivized belief of own absolute score on the test, allowing us to explore how participants update their beliefs.

We find that, conditional on measured ability, women's priors of their own absolute test scores are approximately 0.13 standard deviations of ability lower than men's, a statistically significant gap. Providing highly informative signals of own ability does not significantly reduce this gender gap. After the provision of signals, the gender gap persists at 0.12 standard deviations of ability. Only half of the gender gap in posteriors can be explained by accounting for gender differences in prior beliefs, suggesting significant gender differences in how responsive men and women are to feedback, even given the same prior beliefs. Interestingly, the remaining gender gap is directionally larger in the 90% Signal Accuracy treatment than in the 60% treatment, suggesting that more informative signals of own ability do not more effectively close the gap.

We run a second study to address two major, unresolved questions. First, is the behavior we document well-explained by a Bayesian model? And, second, what is the role for gender stereotypes in predicting responses to feedback? That is, is it the case that women's beliefs are less responsive to feedback generally, or rather, does responsiveness to feedback 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 use a similar experimental paradigm, but expand the range of domains we consider. We test participants 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 (i.e. more male-typed for men, or more female-typed for women).

Participants in the second study complete three rounds of an experiment similar to Study 1. In each round, they take a test of their ability in a randomly-assigned domain. They then provide an incentivized prior belief of both their absolute performance in that test and their belief of their rank relative to others completing the same test. We use the same signal structure as Study 1, but with signal accuracy probabilities of *p=0.5* and *p=0.7*. We then collect posterior beliefs, both of absolute and relative ability. In addition, in Study 2, we collect data not only on their point-wise beliefs of absolute ability (what is your most likely score on the test?), but also on their full prior and posterior distributions over all possible scores. This allows us to address a major question raised by Study 1: how much of gender differences in responses to feedback can be explained by a Bayesian model. Our data in Study 2 allow us to explore gender differences in the shape of belief distributions, and create a Bayesian benchmark for updating behavior for each participant. This framework allows us to parse the results we observe, and uncover systematic departures from the Bayesian model.

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

After the provision of information, stereotypes continue to play a significant role in predicting beliefs. Stereotypes have an impact on posteriors, both through their impact on participant priors (as predicted by the Bayesian model) but also through their impact on how participants react to information given the same prior (a non-Bayesian channel). On average, we find no gender differences in how predictive the Bayesian model is for men and women: the overall pattern is that participants demonstrate conservatism relative to the Bayesian prediction. However, we find 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.

We explore how updating varies depending upon whether the signal drawn was "good" news (a signal greater than or equal to their true score or prior belief) or "bad" news (a signal less than their true score or prior belief). We find that reactions to good and bad news vary with the gender stereotype of the domain. The starkest result is that individuals update less in response to good news in a gender incongruent domain than in a gender congruent 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.

Our paper follows a growing literature on understanding how individuals update beliefs in response to feedback on own ability (see Benjamin 2019 for an extensive overview). Psychology 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 "bad" news but are more likely to attribute "good" news 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 and 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 relatively simple prior belief distributions and compare updating behavior to a full Bayesian benchmark (see, for instance, Mobius et al 2014, 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 (Buser et al 2018, Ertac 2011), consistent with theoretical models of motivated reasoning (such as Rabin and Schrag 1999, Benabou and Tirole 2002, or Koszegi 2006).

This literature has focused on questions of conservatism – are people less responsive to information than the Bayesian model would predict? – and asymmetry – do people respond differently in response to good and 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 2014, Eil and Rao 2011). There is mixed evidence on "good news – bad news" effects. Eil and Rao (2011), Mobius et al (2014), Charness and Dave (2017), and Zimmermann (2020) found participants responded 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. And, 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. In early work, Deaux and Farris (1977) consider how men and women evaluate their own performances on an anagram task that the experimenters have described either as male-typed or female-typed. When the task is described as male-typed, men expect to perform better than women beforehand, and then evaluate their own (known) performance more favorably – in subjective terms - after the fact. They argue that when performance is consistent with expectations – such as men performing well in a male-typed task that they expect to perform well in, or women performing poorly in a male-typed task they expect to perform poorly in -- individuals are more likely to attribute performance to ability, rather than luck. Similarly, in recent work, Shastry, Shurchkov, and Xia (2018) find that negative feedback deters tournament entry for high ability women, primarily because they are too likely to attribute this feedback to ability rather than luck.

There are also studies that attempt to disentangle gender differences within a Bayesian framework. Mobius et al (2014) report that women demonstrate more conservatism than men in their IQ task, updating less in response to information, but that there are no gender differences in "good news – bad news" asymmetry. Coutts (2018) finds very similar results on gender, reporting no gender differences in asymmetry and evidence of more female conservatism (but in both ego-relevant – a math and verbal quiz -- and neutral settings). Ertac (2011) finds mixed results, with women responding less to "good news" than men do about a verbal task, but not about an addition task. Because these studies vary in their paradigms and tasks used, it is hard to know exactly what underlies any across-study differences.

We attempt to unpack the role of stereotypes in driving gender differences in belief updating. We elicit incentivized prior and posterior beliefs of performance, in an environment with uncertainty about true performance. Rather than simply compare men and women in a single domain, and attribute any differences to gender, we compare men and women across a range of domains, and ask how, holding all else fixed, the gender-type of the environment matters for decision-making. To do so, we systematically vary the gender-type of the domain, taking care to hold all other aspects of the environment as fixed as possible. Rather than compare a single verbal task to a single mathematics task, where many features of the task vary, we will compare 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.

In addition, by focusing on absolute ability rather than relative ability, our experiment will yield rich, well-identified data on good and bad news. Men and women across the ability spectrum will be equally likely to receive good or bad news (of equal accuracy). Because participants are updating on absolute ability, the space of possible beliefs will be quite fine, potentially allowing for identification of more subtle differences.[2] 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 (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. STUDY ONE: MOTIVATING EVIDENCE

Design of Study 1

Test of Cognitive Ability

In our 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 are told that they will receive $0.20 for each correct answer if this round of the experiment is selected for payment. Incorrect answers and skipped questions are not penalized.

[2] 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.

Elicitation of Priors

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 would fall 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 distribution over believed performance.

Provision of Signals

Participants are then randomly assigned to one of two signal treatments, either the *60% Signal Treatment* or the *90% Signal Treatment.* We vary *p* in order to study whether more informative feedback is more effective in correcting biased beliefs. 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}.

We explain this mechanism to participants. 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 – 6 (or 1 - 9 for those in the 90% Signal Treatment), the computer will show them their true test score. But, if the computer draws a number between 7 – 10 (or just 10 in the 90% 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, or 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 60% (90%) 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 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 (2019). Full experimental instructions are available in Appendix A.

Implementation

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 available). 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. In Appendix Table B1, we present summary statistics on our participants.

Results of Study 1

Prior Beliefs

There are significant gender differences in performance on the ASVAB test. 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 t-test with

$p<0.001$.[3] On average, participants underestimate their absolute performance on the test when stating their prior beliefs (after performance but before having received a signal). 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) ($p<0.001$). 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 for a detailed treatment of this issue in the context of gender and stereotyping). Figure B1 in the Appendix illustrates this point. 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.

In Table I, we regress prior beliefs of absolute and relative ability on participant gender and performance.[4] Conditional on performance, we estimate that women state beliefs of absolute score approximately 0.6 points lower than men (Column I, $p<0.01$). Gillen, Snowberg, and Yariv (2019) point out that measurement error in control variables can bias coefficients of interest. Here, the main concern is that performance in Round 1 is only 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). In Column II, we add to our model Round 2 performance. We see that our estimated gender gap is unchanged, suggesting that measurement error in performance is not a primary driver of our results. But, we will continue to control for Round 2 score in our specifications for consistency.

In terms of relative ability, women believe they place 7.6 percentage points worse in the ability distribution compared to equally able men (Column III, $p<0.001$). Interestingly, even when we control for a participant's pointwise prior belief of her absolute score, women believe they place worse in the relative distribution than men do (Column IV). 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.

[3] According to the model of stereotyping in Bordalo et al (2016) and as applied in their work on beliefs about gender (Bordalo et al (2019)), this male advantage in performance at the mean may lead to beliefs about self that exaggerate the male advantage on average. Because we study just one domain in Study 1, we cannot directly test for the role of stereotypes in shaping the gender differences we document. This question is instead a central focus of Study 2.

[4] Recall that in this study participants are asked to guess what score they believe they earned. For simplicity, we will refer to this as their "prior", despite the fact that it is a point prediction rather than a distribution.

Table I. Gender Gaps in Prior and Posterior Beliefs of Absolute Ability for Cognitive Skills Test

	OLS Predicting Prior Belief of Absolute Score		**OLS Predicting Prior Belief of Percentile of Score**	
	I	II	III	IV
Female	-0.58*** (0.20)	-0.60*** (0.20)	-0.076**** (0.015)	-0.062**** (0.014)
Round 1 Score	0.60**** (0.022)	0.61**** (0.026)	0.024**** (0.0019)	0.009**** (0.0023)
Round 2 Score		-0.027 (0.030)	-0.003 (0.002)	-0.002 (0.0022)
Prior Belief of Absolute Score				0.024**** (0.0023)
Demographic Controls	Yes	Yes	Yes	Yes
Constant	2.77**** (0.70)	2.83**** (0.70)	0.34**** (0.053)	0.27**** (0.051)
R-squared	0.48	0.48	0.25	0.32
N	981	981	981	981

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 percentile, we assign the participant the midpoint of the percentile bucket she elected when stating her believed rank as the outcome variable in Columns III and IV.

Posterior Beliefs

What happens after participants receive feedback? On average, men revise their beliefs upward by 1.50 points, while women revise their beliefs upward by 1.15 points (p<0.10). Given the signal structure, there are two likely responses: a participant is "convinced" by the signal, and updates her posterior belief to the signal, or a participant is "unconvinced" by the signal, and sticks to her prior belief. (This intuition is made more precise in our theoretical framework in Appendix D.) 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).

In Table II, we regress a participant's posterior belief of absolute score on their gender, their performance, and their signal received. Pooling over the two signal treatments, we see that women's beliefs remain about 0.5 points lower than men's conditional on receiving a noisy signal of performance (Column I). Conditional on having the same performance and after receiving identical signals, women's beliefs are 0.67 points lower than men's in the 60% signal treatment, and 0.42 points lower in the 90% signal treatment (Columns IV, VI, respectively). 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 close the gender gap.

Table II. Gender Differences in Posterior Beliefs of Absolute Ability on Cognitive Skills Test

	OLS Predicting Posterior Belief of Score						
	Pooling Both Signal Treatments			60% Signal Treatment		90% Signal Treatment	
	I	II	III	IV	V	VI	VII
Female	-0.53*** (0.15)	-0.24** (0.12)	-0.074 (0.28)	-0.67*** (0.23)	-0.21 (0.17)	-0.42** (0.19)	-0.32* (0.16)
Round 1 Score	0.42*** (0.053)	0.14*** (0.045)	0.14*** (0.045)	0.42*** (0.058)	0.15*** (0.046)	0.38*** (0.082)	0.17** (0.072)
Round 2 Score	0.005 (0.022)	0.015 (0.018)	0.009 (0.018)	0.007 (0.034)	0.008 (0.026)	0.000 (0.029)	0.021 (0.025)
Signal Received	0.42*** (0.046)	0.38*** (0.037)	0.43**** (0.038)	0.41*** (0.049)	0.38**** (0.037)	0.52**** (0.078)	0.50**** (0.066)
Prior Belief of Rd. 1 Score		0.51**** (0.027)	0.47**** (0.030)		0.51**** (0.027)		0.36**** (0.026)
Female x Signal			-0.12**** (0.031)				
Female x Prior Belief			0.13**** (0.037)				
Demographic Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R-squared	0.78	0.86	0.86	0.74	0.85	0.82	0.87
N	981	981	981	481	481	500	500

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. Columns I – III include a dummy for the assigned signal treatment and the interaction of that signal treatment with signal received and Round 1 score.

In a Bayesian setting, these differences in posterior beliefs (conditional on signal received) must be driven by differences in prior belief distributions. A first step toward exploring the role of prior beliefs is to repeat this analysis while controlling for the pointwise priors that participants report. Pooling over the two signal treatments, we estimate that about half of the posterior gender gap is explained by differences in pointwise priors (Column II). As one might expect, priors play a larger role in the 60% treatment where signals are noisier, explaining roughly 2/3 of the gender gap in posterior beliefs (see Column V). In fact, conditional on measured priors, there is no statistically significant gender gap in posterior beliefs in the 60% treatment. In the 90% signal treatment, priors play a less important role, explaining only approximately 1/4 of the posterior gender gap (Column VII). In the 90% signal treatment, women are significantly less confident than men after receiving a very informative signal of their performance, even conditional on stating the same prior.

Of course, given the design of Study 1, we cannot rule out that differences that remain after controlling for pointwise prior beliefs could be driven by gender differences in prior distributions over all possible scores. That is, despite providing the same guess of their score, a man and a woman could have different distributions –for instance, in terms of variance or skewness - with that same pointwise mode. We return to this issue in Study 2, where we have full data on prior distributions.

Finally, we can ask whether there are gender differences in the extent to which different factors – prior beliefs versus signal received – are predictive of posterior beliefs. We perform this analysis in Column III, including interactions of female with prior and signal received. We estimate that the gender difference in posteriors is driven in part by the fact that women's posteriors are significantly less responsive to signal received and significantly stickier to prior beliefs than men's posteriors are.

Whose Posterior Beliefs are More Accurate?

Overall, signals are quite effective in shifting participant beliefs: most participants report the observed signal as their posterior. If we compute the mean error in beliefs, taking the absolute value of the difference between pointwise belief and true score for each individual, we see that the signals significantly reduce average errors, from 3.27 points in priors to 1.64 points in posteriors ($p<0.001$). Not surprisingly, the reduction is larger in the 90% signal treatment, where mean errors fall from 3.41 points to 1.32 points, than in the 60% signal treatment, (3.13 points to 1.97 points, difference-in-difference significant with $p<0.001$).

Across both signal treatments, men's posterior beliefs are on average significantly closer to the truth than women's are. Conditional on score, signal received, and prior, women's mean errors in posteriors are 0.35 points larger than men's ($p<0.01$). Relative to the accurate beliefs benchmark, posteriors are still too underconfident on average, with participants updating in the correct direction but not enough. Figure B1 in the Appendix provides a visual summary of these results.

Overall, we estimate that the gender gap in beliefs, conditional on performance, falls from 0.60 points to 0.53 points after the provision of signals; we cannot reject that these two gender gaps are the same. Thus, the provision of feedback does not significantly reduce the gender gap in beliefs of own ability. In thinking about the magnitude of these gaps, it may be useful to normalize them against observed performance. The overall gender gap in priors is approximately 0.13 standard deviations of observed ability; this gap is quite similar, at approximately 0.12 standard deviations of ability, in posteriors. Of course, the fact that signals are significantly reducing the mean error in beliefs implies that the gender gap relative to mean error in beliefs is increasing. While the gender gap represents just 18% of the mean error in beliefs in priors, it represents 32% of the mean error in beliefs in posteriors.

Study 1 suggests that signals, while highly effective at reducing mean errors in beliefs, are not particularly effective at closing the gender gap in beliefs of own ability. The results leave open a few important questions. Why do men and women with the same ability, the same prior belief of performance, and the same noisy signal of ability hold different posterior beliefs about their ability? Is it that men and women have different prior *distributions* over possible scores (a potentially Bayesian explanation for the results), or do men and women update differently in response to the news they receive conditional on prior beliefs (a non-Bayesian channel)? And, are these differences a function of gender, or the stereotype of the domain? To provide answers to these questions, we conduct a second experiment aimed at disentangling the role of differences in prior beliefs from differences in updating, and identifying the role of stereotypes in driving the gender differences we observe.

III. STUDY TWO: UNPACKING THE ROLE OF GENDER STEREOTYPES

Design of Study 2

Study 2 builds on the paradigm of Study 1, but (i) explores a variety of domains and (ii) collects additional data on prior and posterior beliefs. Building on the approach 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.[5] While some of these categories lack the external career or educational relevance of the cognitive skills test from Study 1, their clear associated gender-types allow for a better identification of the role of stereotypes in driving beliefs. For each domain, we construct a 20-question multiple-choice test to use as the task. Each multiple-choice test is a timed, 3-minute test, where participants are awarded 1 point for each correct answer. Skipped or incorrect answers are not penalized.

Participants complete three rounds of problem-solving and belief elicitation in the experiment. Each round is structured quite similarly to Experiment 1. First, the participant has three minutes to work on a multiple-choice test from one randomly-chosen domain. We then ask about prior beliefs. Each belief question is incentivized. As in Experiment 1, we ask the participant what they think their most likely score on the test was. But then, we collect additional information geared at better understanding the full distribution of prior beliefs. After reporting what they believe to be their most likely score, on the next page participants are asked the likelihood that they earned this exact score (i.e. “I believe there is a __% chance I earned exactly a score of 6”). After eliciting the probability mass they assign to the mode of their prior, we then ask them for their full distribution over all possible scores, reminding them of the probability mass that they assigned to the mode of their prior. The full instructions for this experiment are available in Appendix A. Finally,

[5] 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 results section.

we ask participants what their believed rank is, comparing themselves to 100 other randomly-chosen participants who completed the same multiple-choice test.[6]

Following the elicitation of priors, we then provide signals of performance, using the language of Study 1. This time, however, we set signal accuracy at either *q=0.5* or *q=0.7*, allowing us to collect more data from participants who receive inaccurate signals of ability. This will be useful in identifying asymmetry in responses to good and bad news. We then re-ask all the beliefs questions, including their believed most likely score, the probability 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 they see a new, randomly-drawn domain.

Following 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 ASVAB questions as a proxy for cognitive skills, 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 Study 2 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. 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%. The study has understanding questions and attention checks in place. 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.

[6] Unlike Study 1, here we allow them to guess any particular rank between 1 – 100. We incentivize them to report the mode of their prior over all possible ranks.

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. 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. Essentially, participants were asked to choose between a lottery that paid out with the probability that they assigned to a specific score, or a lottery that paid out with probability *X.* Participants were asked to provide the *X* that would make them indifferent between the two lotteries. The full language used to explain the procedure is available in Appendix A.

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. This error was corrected in the middle of the experiment, and a comparison of participant answers before and after the error correction finds no evidence that that error impacted the answers given to the questions. A full analysis of this issue is presented in Appendix C.

Econometric Approach

We chose the categories for Study 2 to vary in their associated gender stereotype, as measured by actual and perceived gender gaps in ability. Appendix Table B2 confirms that our categories vary significantly along these dimensions. Figure I illustrates the key data. We arrange the categories by the average slider scale rating given for the category among all participants. 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.[7] The bar graphs illustrate the average male and female score in each domain. 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 even within gender-type.

[7] We will use these classifications when we refer to female-typed or male-typed domains going forward.

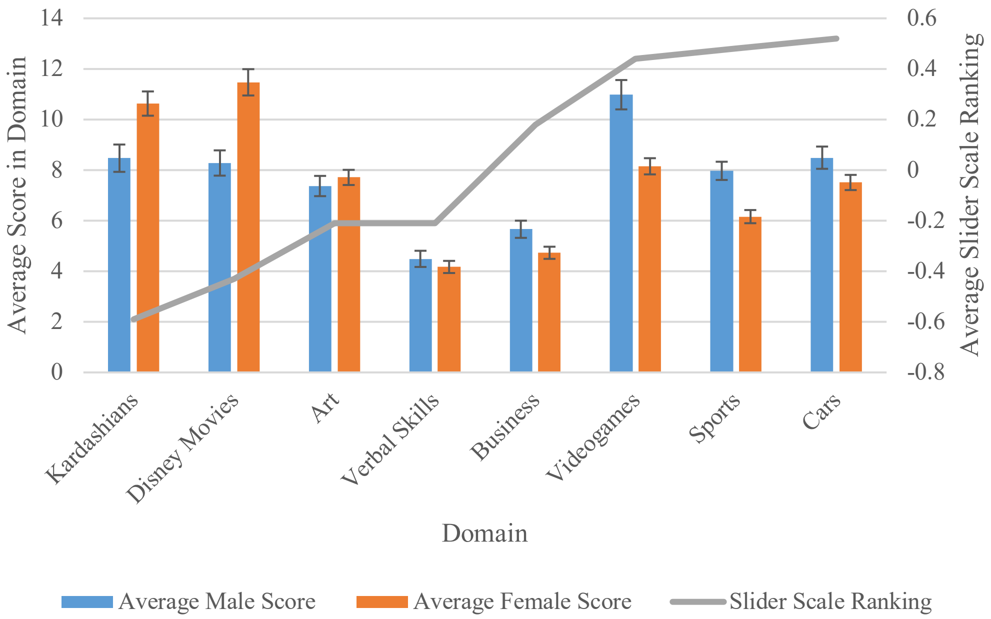

Figure I. Variation in Gender-Type of Domains

Error bars illustrate confidence intervals.

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. Figure B2 in Appendix B presents raw data on overconfidence in priors, to help illustrate two important factors in our data. First, the extent of overconfidence depends largely on a participant's score: higher scoring participants are more underconfident on average, and the relationship is quite linear, consistent with Moore and Healy (2008) and Bordalo et al (2019). While this is a not a focus of our study, it informs our analysis, as we will account for the role of participant score, and task difficulty more generally, in shaping beliefs. The second important factor in predicting overconfidence is the gender-type of the category. While men's overconfidence typically exceeds women's in the male-typed domains for a given score, this is much less consistently the case in the female-typed domains.

A key question is how 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 can test for this type of self-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, holding fixed her own performance.

As in Study 1, measurement error in individual performance is a threat to identification of gender differences and the role of stereotyping. To address this, we again follow the recommendation of Gillen et al (2019). We include in our specifications a second measure of individual performance: the average performance of one's own gender in the category. By including this measure in our regressions, we reduce the chances that the gender gap is informative for beliefs simply because it is proxying for mis-measured individual ability, or gender-specific performance more broadly. The coefficient on the gender gap in our model tells us whether, holding fixed my individual performance *and* the average performance of my gender in the domain, does the relative gender advantage in the domain influence beliefs about myself?

Appendix Table B2 presents summary statistics for our participants in Study 2, as well as an overview of the raw data. We also control for all demographic characteristics in our regressions going forward, including performance on the unincentivized ASVAB questions as a proxy for cognitive skills.

Results for Study 2

Prior versus Posterior Beliefs

We start by exploring the prior beliefs of participants. In Table III, Column I, we explore prior beliefs of absolute ability (see Appendix Table B3 for beliefs of relative ability). First, in Column I, we predict a participant's guess of her most likely score in a category from her gender, her own gender's average advantage in the category (as reported in Panel B of Appendix Table B2), her observed performance, her own gender's average performance in the category, our demographic controls (whether or not she attended high school in the U.S., fixed effects for educational attainment, fixed effects for race, and her score on the ASVAB questions as a proxy of cognitive ability), and round fixed effects.

We estimate that, holding own and gender-specific performance fixed, women report prior beliefs approximately 0.46 points lower than men's for a gender-neutral category (average gender gap in performance of 0). We also see a strong role for the gender congruence of a category in shaping beliefs about self, replicating Bordalo et al (2019). We estimate that a 1-point increase in male advantage in average performance, roughly the size of the male advantage in business, decreases women's beliefs about their own ability by 0.29 points (and, increases men's beliefs about their own ability by 0.29 points). This additional 0.58 points in gender gap is approximately 0.21 SDs of average performance in business. Other similar approaches yield similar results. For instance, if instead of predicting the mode of the participant's prior (her belief of her most likely score), we predicted the mean of her prior belief by computing the weighted average of her distribution over all possible scores, the results are nearly identical.[8]

Table III. Gender Differences and Self-Stereotyping in Prior and Posterior Beliefs

	OLS Predicting Mode of *Prior* Belief of Score	**OLS Predicting Mode of *Posterior* Belief of Score**		
	I	II	III	IV
Female	-0.45**** (0.099)	-0.31**** (0.090)	-0.31**** (0.090)	-0.29**** (0.083)
Own Gender Advantage	0.29**** (0.024)	0.24**** (0.022)	0.23**** (0.029)	0.15**** (0.021)
Score	0.61**** (0.013)	0.54**** (0.026)	0.54**** (0.026)	0.39**** (0.017)
Signal 70% Accurate Treatment x Own Gender Advantage			0.00 (0.038)	
Bayesian Predicted Posterior Belief				0.45**** (0.016)
Demographic Controls	Yes	Yes	Yes	Yes
R-squared	0.45	0.64	0.64	0.68
Clusters (Obs.)	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. In Columns II and III, we also control for signal received, signal treatment, and the interactions of signal treatment with signal received and score.

In Appendix Table B5, we offer analysis of the shape of the prior distributions provided by participants, with a focus on variance and skewness. 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

[8] See Appendix Table B4 for these results. There, we also present results when we take 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.

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. See Appendix Table B5 for more details.

How do participant beliefs respond to the signals received? Figure II provides some initial evidence.[9] In Figure II Panel (a), we graph the average change in beliefs (mode of posterior – mode of prior), splitting 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. The extent of this movement varies. In the male-typed categories, men's beliefs increase directionally more than women's, while in the female-typed categories, women's beliefs increase significantly more than men's.

In Panel (b), we focus on the "type" of posterior report. The two most frequent types of reports are (i) participants reporting the mode of their prior as the mode of their posterior, and (ii) participants reporting the signal as the mode of their posterior. Essentially, participants see the feedback and either stick to what they had believed initially (reporting the mode of their prior as their posterior), or, they update fully (reporting the signal as their posterior). Again, these types of report vary by gender and gender-type of category. In male-typed categories, women are more likely to stick to their prior beliefs than men; this pattern reverses in female-typed categories. Men are more likely to report the signal as the mode of their posterior than women are in both types of categories, but this gender gap is larger in male-typed categories. In the analysis below, we consider more formally the role for stereotypes in driving these patterns.

Do signals reduce the gender gap in beliefs, and in particular the reliance on stereotypes in shaping beliefs about own ability? Table III, Column II replicates Column I but instead predicts posterior beliefs. Directionally, the estimated gender gap for a gender-neutral category and the reliance on stereotypes shrink in posterior beliefs compared to prior beliefs. However, both remain 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 shift in own gender advantage increases beliefs of own ability in posteriors by 0.24 points (see Column II, effect of own gender advantage on posteriors is significantly different than 0 with $p<0.01$).[10]

In Column III, we ask whether more informative signals are any 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

[9] Note that the vast 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. Only 159 signals fall outside of the 0 – 20 range. For completeness, we include all observations in our main text analysis.

[10] 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. See Appendix Table B4.

accuracy treatment? The answer is no, as the estimated interaction is very close to 0. As in Study 1, more informative feedback is no more effective in reducing gender gaps. Given the similarity of the results across the two signal treatments, we will consider them jointly in our remaining analysis.

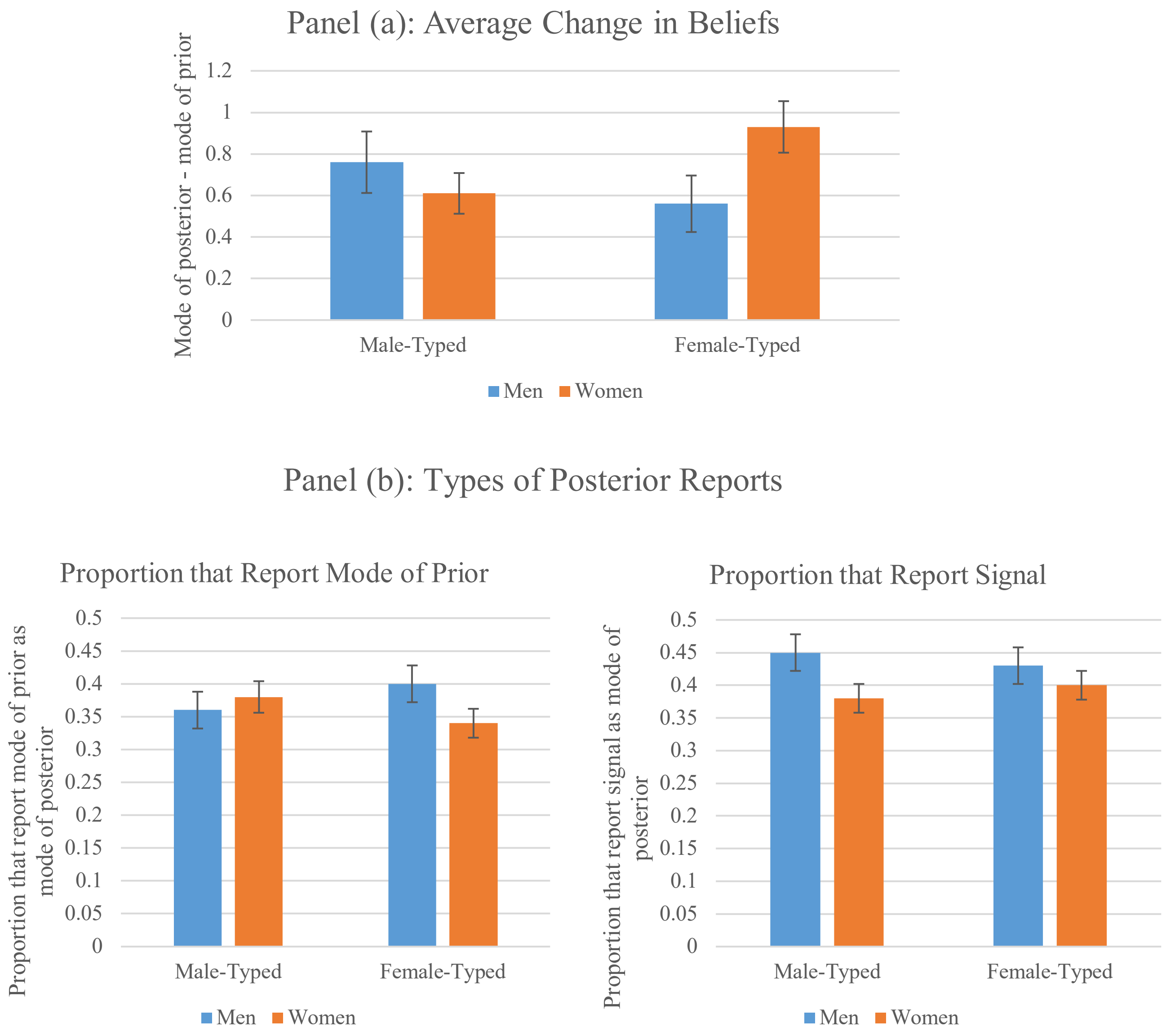

Figure II. Overview of Evidence on Posterior Beliefs

Error bars illustrate 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).

Figure III compares overconfidence in prior and posterior beliefs. We graph the linear fit of absolute overconfidence (believed score – observed score) against true score for both men and women, graphing priors in Panel (a) and posteriors in Panel (b). Within each panel, we split the sample by whether the observation is from a male-typed or female-typed domain. This presentation of the data has the advantage of allowing us to compare estimates of overconfidence for men and women with the same score by comparing the height of the fitted line at different points in the score distribution. Essentially, we are interested in the vertical distance between the lines for a given score.[11,12]

Perhaps the most striking result from both panels is the high degree of similarity in observed overconfidence of *men in male-typed domains* and *women in female-typed domains*. If we compare men and women's beliefs in domains that are gender congruent, there are no gender differences on average. The lines essentially lie on top of each other. Rather, we see more underconfidence from women than men only within male-typed domains. When we focus on female-typed domains, women are actually less underconfident than men for a wide range of scores. Turning our attention to the comparison of priors and posteriors, we can clearly see that while beliefs become significantly more accurate on average, the gender gaps in posteriors are quite similar in size when compared to prior beliefs.

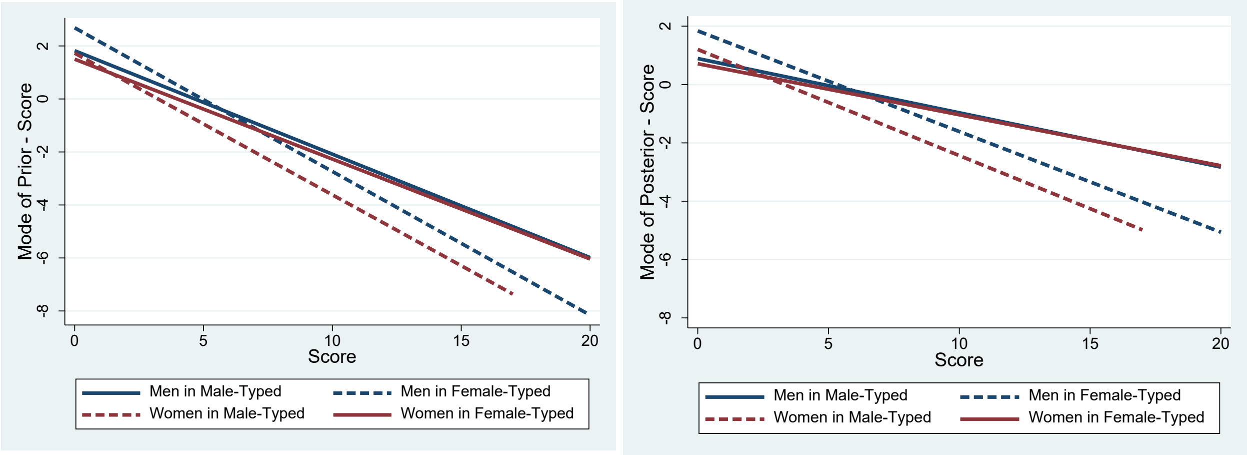

Figure III. Overconfidence in Prior versus Posterior Beliefs

Panel (a) graphs prior beliefs and Panel (b) graphs posterior beliefs. The lines represent linear fits of observed overconfidence on 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).

[11] As a point of reference, the average score is 7.6 (median of 7).

[12] We are not particularly interested in the slopes of each line, as these simply tell us how the degree of overconfidence varies with score. But, we know that score is a key factor for the degree of overconfidence (see Moore and Healy 2008 for example). So, it is important to control for score, which we do throughout our analysis and our visual presentation of the data.

Are Gender Differences in Posteriors Well-Explained by the Bayesian Model?

Are these remaining gender differences well-explained by the Bayesian model? We saw in Study 1 that gender gaps in posteriors were only partially explained by differences in prior guesses of most likely score across men and women. Our data in Study 2 are much more extensive, allowing us to provide a fuller test of the Bayesian model. In particular, for each participant, we can use her reported prior distribution and the signal she received to make a Bayesian prediction for what the reported mode of her posterior should be. This Bayesian prediction tells us, given the signal structure, the full distribution of prior beliefs reported by the participant, including the probability mass she assigns to each possible score, and the signal she receives, what score a Bayesian decision-maker would report as the mode of their posterior. We do this for each person-round observation in our dataset.

In Appendix D, we present a simple Bayesian model that generates these predictions. The key takeaway is that, in our framework, the Bayesian prediction for most participants is of one of two types: the mode of her posterior should be the signal she observed, or the mode of her posterior should be the mode of the prior she reported. 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 distribution, she should update to that signal after observing it. If she did not put sufficient weight on the signal ex ante, then she should continue to report the mode of her prior as the mode of her posterior. 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. These participants have essentially seen a probability zero 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.[13]

Turning to the rest of participants, for 51% of observations from men and 45% of observations from women, the Bayesian prediction is that the participant should report her signal. For 51% of observations from men and 56% of observations from women, the Bayesian prediction is that the participant should report the mode of her prior as the mode of her posterior. Finally, for 10% of observations from men and 11% of observations from women, the Bayesian prediction is some other score (not the signal nor the mode of the

[13] Another focal response would be for these participants to report the mode of their prior, however it seems odd to classify this as a Bayesian response, since the signal they have seen directly implies that the mode of their prior has zero probability of being their true score.

prior) – 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).[14] Note that these proportions do not sum to 100% because for some participants, the signal is the mode of their prior.

The question we want to ask is whether the posteriors we observe are well-explained by the Bayesian model. To answer this question, we can predict a participant's observed posterior from the Bayesian prediction for her posterior, and we can ask whether, conditional on these Bayesian predictions, there is an additional impact of gender or the gender-type of the category. In Table III, Column IV, we add the Bayesian prediction for a participant's posterior as a control. If the impact of gender and stereotyping were well-explained by the Bayesian model, we would expect reduced coefficients on gender and own gender advantage as compared to Columns II and III. In particular, if the Bayesian model explains the impact of gender and own gender advantage on posterior beliefs, we should see a large reduction in the coefficients on these explanatory variables when comparing Column II and Column IV.

While the Bayesian prediction does have significant predictive power for observed posteriors, we see systematic departures from Bayes that are well-predicted by gender and gender stereotypes. That is, we estimate a significant role for both gender and stereotypes *on top of* the role that the Bayesian model predicts. First, we estimate a significant gender gap in posterior beliefs for gender-neutral categories, conditional on the Bayesian prediction. We estimate that women report posteriors roughly 0.3 points worse than men for a gender-neutral category, conditional on having the same Bayesian-predicted posterior. And, we estimate a significant impact of stereotypes in excess of what the Bayesian model would predict. 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. That is, even after accounting for how stereotypes would impact posteriors in a Bayesian model, we estimate that stereotypes have significant predictive power for posterior beliefs. Stereotypes seem to color the way participants update in response to information.[15]

Overall, we replicate past findings of conservatism (or, in the language of Benjamin (2019), under-inference from the signals). If we regress posterior beliefs simply on Bayesian predictions, the slope is significantly

[14] 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.

[15] In Appendix Table B6, Column I, 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. In Appendix Table B6, Column II, we repeat this analysis but excluding any individual who observed a "zero probability event", having put no prior weight on any score that could have generated the signal. The Bayesian prediction is more predictive in this model, but there continues to be a significant, albeit smaller, impact of own gender advantage (estimated at 0.07 (SE 0.019)). It's worth noting that this analysis is a rather demanding test of the hypothesis, as it excludes those individuals whose prior beliefs were most distorted by stereotypes.

less than one. For both men and women, posteriors are more responsive to the Bayesian prediction when the domain is gender congruent. We formalize this finding in Table IV, predicting posterior beliefs from the Bayesian predictions and splitting the sample by male-typed and female-typed categories.

Table IV. Stereotypes Explain Departures from Bayesian Predictions

OLS Predicting Posterior Belief of Absolute Score					
	Male-Typed Domains		Female-Typed Domains		Overall
	I	II	III	IV	V
Female	-0.79**** (0.107)	-0.06 (0.191)	0.16 (0.111)	-0.59*** (0.229)	-0.40** (0.159)
Bayesian Prediction	0.44**** (0.022)	0.50**** (0.027)	0.46**** (0.024)	0.38**** (0.035)	0.44**** (0.021)
Female x Bayesian Prediction		-0.11**** (0.030)		0.11**** (0.032)	0.01 (0.020)
Own Gender Advantage					-0.04 (0.036)
Own Gender Advantage x Bayesian Prediction					0.03**** (0.004)
Score	0.37**** (0.022)	0.36**** (0.022)	0.41**** (0.024)	0.41**** (0.024)	0.38**** (0.017)
Demographic Controls	Yes	Yes	Yes	Yes	Yes
R-squared	0.64	0.64	0.71	0.71	0.69
Clusters (Obs.)	1832 (2992)	1832 (2992)	1839 (2975)	1839 (2975)	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. 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 estimate that women hold significantly lower beliefs than men conditional on the Bayesian prediction in male-typed categories (Column I); in these domains, women's beliefs are on average less responsive to the Bayesian prediction (Column II). This pattern is reversed in female-typed domains, where men's posterior beliefs are less responsive to the Bayesian prediction than women's (Columns III, IV). In Column V, we produce the interacted model. We see that for an estimated gender-neutral category, there are *no* gender differences in how predictive the Bayesian prediction is for posterior beliefs. But, own gender advantage plays an important role. The Bayesian prediction better predicts posterior beliefs in gender congruent categories.

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 differences, 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), the relative roles of these two biases is clear in our setting: under-inference is the relatively larger deviation from Bayesian updating.

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 idea 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 E.

Past literature has also 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, exactly the opposite pattern.

Beyond the Bayesian Model

From a theoretical perspective, the Bayesian prediction incorporates all information about the participant prior that should be relevant for the mode of the posterior reported. In practice, however, different priors (and in particular modes/means of priors), despite generating the same Bayesian prediction, might produce different expected posteriors in a more intuitive (not Bayesian) model. For example, consider two participants both with a true score of 9 in a category. Suppose Participant A reports a mode of her prior of "7", while Participant B reports a mode of her prior 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 the mode of their posterior (assuming Participant A puts sufficient weight on the possibility of her score being 9 in her prior). 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 the mode of her posterior than Participant A would be, as the signal is more in line with Participant B's prior belief.

Our previous analysis asked – given the same Bayesian prediction for behavior – do men and women vary in their posterior beliefs. We could ask a slightly different question, which is, conditional on what we observe about participant priors – in particular, the mode of the prior, the prior weight assigned to it, the prior weight assigned to the signal, the standard deviation of the prior, and the degree of positive skewness – do men and women vary in their posterior beliefs? Do stereotypes have predictive power conditional on having not just the same Bayesian prediction, but also the same salient features of priors?

We report the results in Appendix Table B7. Including the 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 eliminates the estimated gender gap in posterior beliefs for a gender-neutral category; however, a significant, albeit smaller (approximately 1/3 the size), role for stereotypes persists. It continues to be the case that, conditional on both the Bayesian prediction and more detailed information about participant priors, own gender advantage shapes posterior beliefs.

Good News, Bad News

One interesting question that our framework allows us to consider is asymmetry in updating in response to "good" or "bad" news. Past studies, including Eil and Rao (2011) and Mobius et al (2014), have found that participants respond more to good news than bad news (relative to what the Bayesian model would predict) when updating on own ability. With our data, we can explore whether the extent of asymmetry varies with gender or gender-type of the domain.

To explore this question, we first define what it means to receive good or bad news. In our primary analysis, we will refer to a signal as good news if it is equal to or above their true score, and we will refer to a signal as bad news if it is below their true score.[16] 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 underconfident 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 the mode or mean of their prior. In this way, our analysis is like an "intent-to-treat". In Appendix Table B10, we perform the same analysis but instead define good and bad news relative to priors and find, in general, quite similar results.

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 more than women. In female-typed domains, this pattern is exactly reversed: women change their beliefs significantly more than men in response to good news in female-typed domains. Similarly, reactions to bad news also seem to depend on the gender-type of the domain. 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. This provides suggestive evidence that asymmetry is indeed a function of the gender congruence of the domain.

To formalize this, we follow the econometric approach of Eil and Rao (2011). Suppose we predict posterior beliefs from the Bayesian prediction of posteriors. A good news or bad news effect could take two different forms. First, we could simply include a dummy to indicate that the news received was good (signal greater than or equal to score). A positive coefficient on this dummy would indicate what Eil and Rao (2011) refer to as a "generalized optimism" – beliefs that are on average greater than what the Bayesian model would predict for this particular type of news relative to when the same Bayesian prediction is made for bad news.

[16] 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 (61% of cases) then below a participant's prior (26% of cases).

We can also test for differential responsiveness to the Bayesian prediction for good news, indicated by a different slope on the Bayesian prediction depending upon whether the news was good. A steeper slope – indicated by a positive interaction term on the good news dummy and the Bayesian prediction – suggests greater responsiveness to the Bayesian prediction.

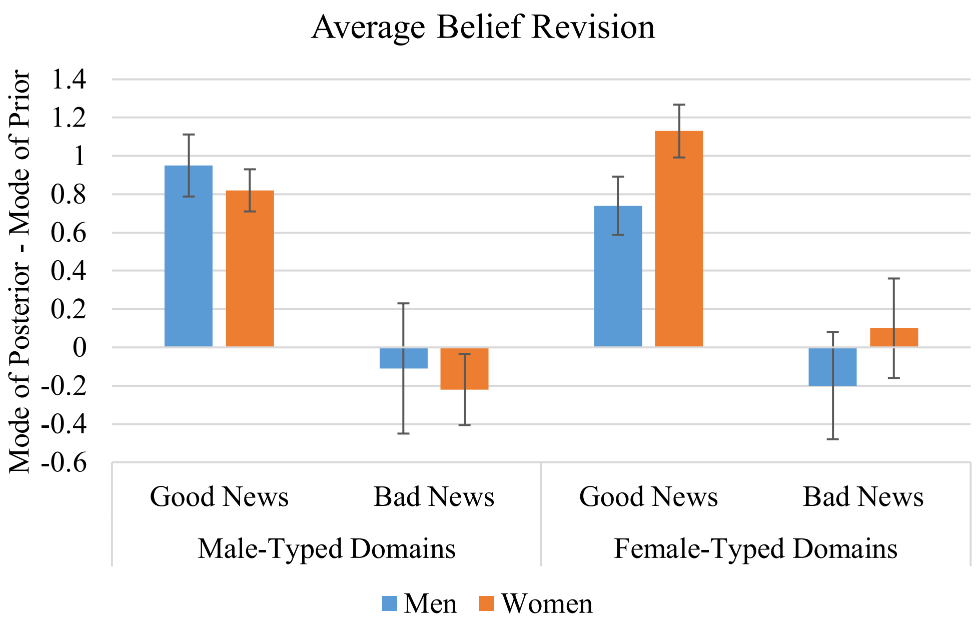

Figure IV. Average Belief Revisions after Good and Bad News

Error bars illustrate 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).

To simplify interpretations, we run a very basic model: we predict a participant's reported posterior belief from the Bayesian prediction for her posterior, a dummy for whether she received good news, and the interaction of the two. We leave out all other explanatory variables so that we can interpret and compare constants across models simply. We split the data according to gender-type of the domain and gender. The results are presented in Table V, Panel A.

Let's start by considering male-typed domains. 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, men have a larger estimated intercept for bad news than women do (a constant of 2.11 as opposed to 1.43). This suggests that after receiving bad news, men report higher posterior beliefs than women conditional on having the same Bayesian prediction. Turning to good news in male-typed domains, men are estimated to be much more responsive than women to the Bayesian prediction (estimated coefficient of 0.79 for men and 0.59 for women, p<0.01 in interacted model, see Appendix Table B9). It seems to be the case that women discount increasingly good news more than men do when it comes in a male-typed domain.[17] So, in male-typed domains, we estimate that the gender gap in posteriors after bad news is sizable, and flat over the range of Bayesian predictions. The gender gap in posteriors after good news in male-typed domains grows with the Bayesian prediction, eventually exceeding the gender gap in posteriors after bad news.

These patterns look quite different when we turn to female-typed domains, suggesting that much of the gender difference in responsiveness to good and bad news is a function of gender stereotypes. When we consider participants who receive bad news in a female-typed domain, women are more responsive than men (coefficient of 0.92 versus 0.79, p<0.01). We also estimate a smaller constant for women than men after receiving bad news (0.90 versus 1.34). This suggests that among participants with a very low Bayesian prediction, men will display greater overconfidence relative to that prediction than women; but, as the Bayesian prediction increases, given women's greater responsiveness, women will ultimately end up with higher beliefs conditional on the Bayesian prediction. After receiving good news, women are also much more responsive than men, exactly 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, see interacted model in Appendix Table B8). We also estimate a smaller constant for women than men after receiving good news, again suggesting that for participants for whom the Bayesian prediction is quite close to 0, men's posterior beliefs will exceed the Bayesian prediction by more than women's do; however, given the differences in responsiveness, this gender gap reverses as the Bayesian prediction increases.

An interesting consequence of these patterns is that we observe substantial differences in the predictive power of the Bayesian model by gender stereotype. When individuals are operating in a gender congruent domain, the model r-squareds are rather high (0.64 for men in male-typed domains, 0.73 for women in female-typed domains). But, when individuals are operating in gender incongruent domains, posteriors are much harder to predict (r-squared of 0.47 for women in male-typed domains, r-squared of 0.53 for men in female-typed domains). The Bayesian model has much more predictive power in gender congruent domains than gender incongruent domains.

[17] The estimated constant for women receiving good news is larger than the estimated constant for men (2.08 compared to 1.57), suggesting that for those whose Bayesian prediction is quite close to 0 women will be directionally more overconfident than men after receiving good news. However, given the differences in responsiveness, this pattern reverses quite quickly as the Bayesian prediction increases.

Table V. Good News and Bad News

Panel A: Regression Output

	OLS Predicting Posterior Belief			
	Male-Typed Domains		Female-Typed Domains	
	Men	Women	Men	Women
Bayesian Prediction	0.76**** (0.052)	0.75**** (0.040)	0.79**** (0.055)	0.92**** (0.022)
Good News Dummy	-0.54 (0.44)	0.65** (0.25)	0.77* (0.45)	0.46* (0.24)
Good News x Bayesian Prediction	0.030 (0.057)	-0.16**** (0.048)	-0.11* (0.066)	-0.12**** (0.028)
Constant	2.11**** (0.41)	1.43**** (0.21)	1.34**** (0.37)	0.90**** (0.21)
R-squared	0.64	0.47	0.53	0.73
Cluster (Obs.)	735 (1213)	1097 (1779)	716 (1139)	1123 (1836)
Estimated Responsiveness to Good News	0.79	0.59	0.68	0.80
Estimated Responsiveness to Bad News	0.76	0.75	0.79	0.92

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).

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

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 much 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. Thus, stereotypes for understanding differences in responsiveness to good news.

It is worth noting that the most challenging task would seem to be convincing individuals of their talent in gender incongruent domains. Across the four cells, individuals are least responsive to good news in gender incongruent domains. Men are least responsive to good news in female-typed domains. Similarly, women are least responsive to good news in male-typed domains. Responsiveness is lowest when individuals receive better than expected information in a gender incongruent domain.

In Appendix Table B9, we replicate these results using only the sub-sample of participants who put some positive prior probability on at least one score that could have generated the observed signal (the excluded participants are those who essentially observed an event that they assigned zero probability to in their prior). This ensures that the Bayesian prediction is clear for every participant in the sub-sample; however, it comes at the important cost of excluding many of those participants whose prior beliefs are most biased by stereotypes. Among this sub-sample, the Bayesian predictions are in general much more predictive of posterior beliefs. But, we continue to find that individuals (both men and women) are more responsive to good news when it comes in a gender congruent domain than when it comes in a gender incongruent domain. However, we find weaker evidence that women are more responsive to bad news than good, suggesting this result may be driven in part by people whose initial beliefs are quite biased, and who subsequently do not react to "large" good surprises.

In Appendix Table B10, 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 in general quite similar results. In particular, responsiveness to good news is larger when it arrives in a gender congruent than in a gender incongruent domain.

Finally, in Appendix E, we re-revisit these results in the alternative framework of the Grether (1980) model. Again, we find support for our main finding of interest here. In particular, 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.

IV. CONCLUSION

There is increasing evidence that stereotyped beliefs drive 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 own ability absent feedback. We then show that self-stereotyping is also highly predictive of beliefs after feedback. This operates through two main channels: a Bayesian channel through which stereotypical prior beliefs fuel stereotypical posterior beliefs, and a non-Bayesian channel through which stereotyping shapes updating behavior. In particular, we see that individuals deviate from the Bayesian model, and these deviations are explained in part by gender stereotypes. Holding fixed the Bayesian prediction for beliefs, individuals hold more optimistic beliefs about their performance as the domain becomes more gender congruent. Within both of our experimental settings, more informative feedback is no more effective in closing the gender gap or reducing reliance on gender stereotypes.

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.

Our work advances our understanding of the ways in which individuals deviate from the Bayesian model in updating their beliefs. We see that individuals are more conservative, or under-infer more from signals, particularly good news signals, in gender incongruent domains. These differences by gender-type of the domain are not 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 this full set of results 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 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 large part on whether participant priors are over or underconfident on average.

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 on the part of 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 pervasiveness and power of self-stereotyping. Stereotypes do not just impact beliefs about ability when information is scarce; rather, it appears stereotypes color 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. 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. 2016. “Stereotypes.” *The Quarterly Journal of Economics* 131 (4): 1753–94. https://doi.org/10.1093/qje/qjw029.

Born, Andreas, Eva Ranehill, and Anna Sandberg. 2020. “Gender and Willingness to Lead – Does the Gender Composition of Teams Matter?” *Working Paper*. https://doi.org/10.2139/ssrn.3207198.

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. 2019. “When to Apply.” *Working Paper*.

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.

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 and Judd Kessler. 2019. "The Gender Gap in Self-Promotion." *Working paper*.

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. 2014. “Managing Self-Confidence.” *Working Paper*, August.

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

Experiment Materials (under separate cover)

APPENDIX B

In Table B1, we present summary statistics for participants in Study 1.

Table B1. Summary Statistics for Study 1

	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	

Table B2. Summary Statistics for Study 2

	Men	Women	P-value from test of proportions
White	0.79	0.82	0.12
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.

	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 B3. Gender Differences in Priors and Posteriors of Relative Ability

	OLS Predicting Prior Belief of Relative Rank (1=Best; 100=worst)		**OLS Predicting Posterior Belief of Relative Rank (1 = Best, 100 = Worst)**	
	I	II	III	IV
Female	4.62**** (0.872)	3.53**** (0.847)	5.27**** (0.849)	4.47**** (0.828)
Own Gender Advantage	-1.71**** (0.205)	-1.00**** (0.201)	-1.64**** (0.199)	-1.03**** (0.198)
Score	-2.20**** (0.114)	-0.74**** (0.145)	-1.80**** (0.235)	-0.41* (0.242)
Belief of Absolute Score (Prior in Columns I/II; Posterior in Columns III/IV)		-2.41**** (0.158)		-2.56**** (0.176)
Demographic Controls	Yes	Yes	Yes	Yes
R-squared	0.13	0.19	0.17	0.23
Clusters (Obs.)	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. We also control for signal received, signal treatment, and the interactions of signal treatment with signal received and score in Columns III and IV.

Table B4. Robustness Analysis of Prior and Posterior Beliefs

Panel A: Prior Beliefs				
	OLS Predicting Belief of Score: Mode of Prior	**OLS Predicting Belief of Score: Mean of Prior**	**OLS Predicting Belief of Score: Mode of Prior**	**OLS Predicting Belief of Score: Mean of Prior**
	I	II	III	IV
Female	-0.45**** (0.099)	-0.38**** (0.105)	-0.48**** (0.099)	-0.41**** (0.105)
Own Gender Advantage in Performance	0.29**** (0.024)	0.30**** (0.023)		
Own Gender Advantage in Perception			1.18**** (0.097)	1.22**** (0.102)
Score	0.61**** (0.013)	0.62**** (0.013)	0.61**** (0.013)	0.62**** (0.014)
Demographic Controls	Yes	Yes	Yes	Yes
R-squared	0.45	0.43	0.44	0.43
Clusters (Obs.)	1989 (5967)	1989 (5967)	1989 (5967)	1989 (5967)

Panel B: Posterior Beliefs				
	OLS Predicting Belief of Score: Mode of Posterior	**OLS Predicting Belief of Score: Mean of Posterior**	**OLS Predicting Belief of Score: Mode of Posterior**	**OLS Predicting Belief of Score: Mean of Posterior**
	I	II	III	IV
Female	-0.31**** (0.090)	-0.36**** (0.096)	-0.33**** (0.089)	-0.38**** (0.096)
Own Gender Advantage in Performance	0.24**** (0.022)	0.24**** (0.021)		
Own Gender Advantage in Perception			1.00**** (0.089)	0.99**** (0.091)
Score	0.54**** (0.026)	0.55**** (0.027)	0.54**** (0.026)	0.55**** (0.027)
Demographic Controls	Yes	Yes	Yes	Yes
R-squared	0.64	0.60	0.64	0.60
Clusters (Obs.)	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. We also control for signal received, signal treatment, the interaction of signal treatment with signal received and score in Panel B.

Analysis of Gender Differences and Stereotypes in Predicting Shapes of Priors

After eliciting a participant's belief of her most likely score in the category, we then elicit her beliefs over all possible scores, gathering a complete picture of her prior distribution. In this sub-section, we explore the properties and shapes of those priors. For simplicity we focus on two key distributional measures: variance and skewness. In general, participant priors are quite narrow. 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 Figure B3 for a histogram of the ranges). The average standard deviation of a prior distribution in our sample is 1.39 points. We define three buckets of skewness in priors. 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 finally a right-skewed bucket in which the mean exceeds the median. Approximately 21% of distributions are symmetric under this definition, 39% are left-skewed, and 40% are right-skewed. On average, the mean of a participant's prior is approximately 0.05 points greater than the median of her prior.

Table B5 presents comparisons by gender and gender stereotype, all conditional on score and demographics. Notably, we see significant and rather consistent gender differences in terms of both variance and skewness. Women report tighter, lower variance priors than men on average. As can be seen in Column I of Table B5, women assign directionally more probability mass to the mode of their prior. And, as seen in Columns II and III, women's priors have significantly smaller ranges and less variance. In Columns IV and V, we see that women are significantly less likely to report a symmetric prior, and have a slightly larger degree of right-skewness on average. These differences are independent of the gender-type of the category.

There are fewer and less consistent differences by gender stereotype. First, we see no relationship between the degree of skewness in the prior and the extent to which a category is gender congruent (Columns IV and V). The relationship between gender congruence and variance is also weak. On the one hand, participants place significantly more probability mass on the mode of their prior in more gender congruent categories ($p<0.01$). This would seem to be indicative of less uncertainty in their prior for more gender congruent categories. But, on the other hand, neither the range nor the standard deviation of the prior decrease with gender congruence. Thus, the only clear takeaway in terms of gender stereotypes would seem to be that participants feel more sure about the mode of the prior for more gender congruent categories. 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 B5. Gender Differences and Stereotypes in Shapes of Priors

	OLS Predicting Mass 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	1.46* (0.848)	-0.62**** (0.184)	-0.15*** (0.050)	-0.03** (0.013)	0.03* (0.016)
Own Gender Advantage	0.53*** (0.182)	0.00 (0.033)	0.01 (0.009)	0.00 (0.003)	-0.01 (0.004)
Score	0.23** (0.097)	-0.01 (0.020)	-0.01* (0.006)	-0.00*** (0.002)	0.00 (0.002)
Demographic Controls	Yes	Yes	Yes	Yes	Yes
R-squared	0.03	0.06	0.06	0.02	0.01
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 B6. Robustness Analysis of Bayesian Predictions versus Observed Posteriors

	OLS Predicting *Mean* of Posterior Belief of Score	**OLS Predicting Mode of Posterior Belief of Score Excluding "Zero Probability Events"**
	I	II
Female	-0.30**** (0.086)	-0.21*** (0.073)
Own Gender Advantage	0.15**** (0.020)	0.07**** (0.019)
Score	0.32**** (0.021)	0.27**** (0.015)
Bayesian Prediction of Mean	0.52**** (0.019)	0.67**** (0.014)
Demographic Controls	Yes	Yes
R-squared	0.67	0.79
Clusters (Obs.)	1989 (5967)	1982 (5420)

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. In Column I, the mean of the Bayesian-predicted posterior is computed as the signal for participants who put no positive probability on any score that could have generated the signal in their prior distribution. Column II excludes any observation in which an individual put zero prior probability mass on every score that could have generated the observed signal.

Table B7. Using Other Characteristics of the Prior to Predict Posteriors

OLS Predicting Observed Posterior		
	I	II
Female	-0.04 (0.056)	-0.13 (0.101)
Bayesian Prediction	0.03 (0.019)	0.02 (0.022)
Female x Bayesian Prediction		0.01 (0.014)
Own Gender Advantage	0.05*** (0.016)	-0.05** (0.026)
Own Gender Advantage x Bayesian Prediction		0.01**** (0.003)
Score	0.39**** (0.015)	0.39**** (0.015)
Mode of Prior	0.62**** (0.020)	0.62**** (0.019)
Weight on Mode of Prior	0.00** (0.001)	0.00* (0.001)
Weight on Signal in Prior	0.00 (0.001)	0.00 (0.001)
SD of Prior	0.10*** (0.032)	0.10*** (0.032)
Right-skewness of Prior	0.27**** (0.066)	0.27**** (0.066)
Noise Draw	0.24**** (0.015)	0.24**** (0.015)
Demographic Controls	Yes	Yes
R-squared	0.82	0.82
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. Noise draw is the random draw of an integer drawn from [-5,5] that was added to the score to determine the signal observed. 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 B8. Good News and Bad News Interacted Models

	OLS Predicting Posterior Belief of Score	
	Male-Typed Domains	Female-Typed Domains
Bayesian Prediction	0.76**** (0.052)	0.79**** (0.053)
Good News Dummy	-0.54 (0.436)	0.77* (0.43)
Good News x Bayesian Prediction	0.03 (0.057)	-0.11* (0.063)
Female	-0.69* (0.462)	-0.44 (0.41)
Female x Bayesian Prediction	0.02 (0.065)	0.13** (0.057)
Female x Good News	1.19*** (0.504)	-0.31 (0.49)
Female x Good News x Bayesian Prediction	-0.19*** (0.074)	-0.01 (0.069)
Constant	2.11**** (0.409)	1.34**** (0.35)
R-squared	0.60	0.68
Cluster (Obs.)	1832 (2992)	1839 (2975)

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.

Table B9. Good News and Bad News Excluding "Zero Probability Events"

	OLS Predicting Posterior Belief of Score			
	Male-Typed Domains		Female-Typed Domains	
	Men	Women	Men	Women
Bayesian Prediction	0.80**** (0.051)	0.82**** (0.028)	0.81**** (0.054)	0.97**** (0.019)
Good News Dummy	-0.83* (0.44)	0.18 (0.189)	0.3 (0.45)	0.44** (0.209)
Good News x Bayesian Prediction	0.13** (0.053)	0.002 (0.035)	0.01 (0.064)	-0.05** (0.023)
Constant	1.72**** (0.42)	0.99**** (0.15)	1.17**** (0.38)	0.41**** (0.18)
R-squared	0.74	0.67	0.62	0.84
Cluster (Obs.)	700 (1106)	1040 (1617)	689 (1057)	1060 (1640)
Estimated Responsiveness to Good News	0.93	0.82	0.81	0.93
Estimated Responsiveness to Bad News	0.80	0.82	0.81	0.97

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. Excludes any observation in which an individual put zero prior probability mass on every score that could have generated the observed signal.

Table B10. Good News and Bad News Defined Relative to Priors

	OLS Predicting Posterior Belief			
	Male-Typed Domains		Female-Typed Domains	
	Men	Women	Men	Women
Bayesian Prediction	0.79**** (0.035)	0.81**** (0.031)	0.73**** (0.06)	0.9**** (0.021)
Good News Dummy	-0.65* (0.366)	1.1**** (0.23)	-0.16 (0.505)	0.36 (0.222)
Good News x Bayesian Prediction	0.004 (0.043)	-0.27**** (0.041)	-0.03 (0.069)	-0.11**** (0.028)
Constant	2.10**** (0.328)	1.18**** (0.188)	2.01**** (0.469)	1.01**** (0.187)
R-squared	0.64	0.48	0.52	0.73
Cluster (Obs.)	735 (1213)	1097 (1779)	716 (1139)	1123 (1836)
Estimated Responsiveness to Good News	0.79	0.54	0.70	0.79
Estimated Responsiveness to Bad News	0.79	0.81	0.73	0.90

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 mode of prior.

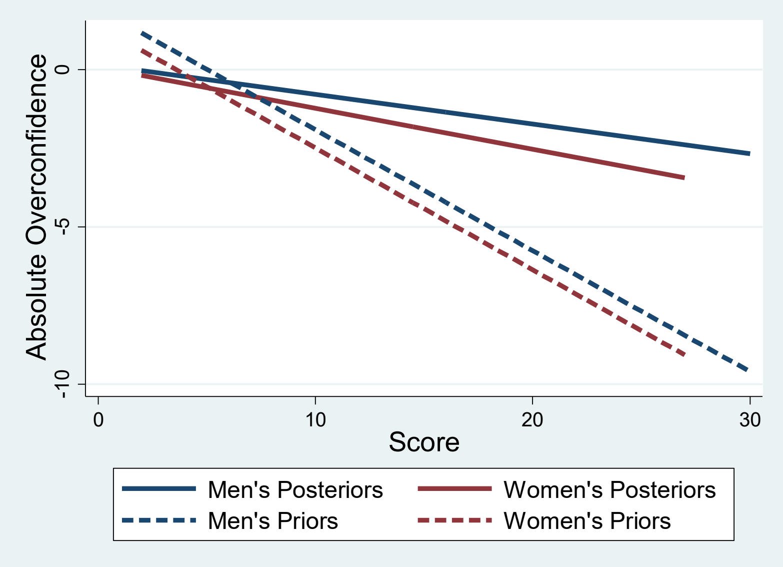

Figure B1. Linear Fits of Overconfidence (in Absolute Score) in Priors and Posteriors in Study 1

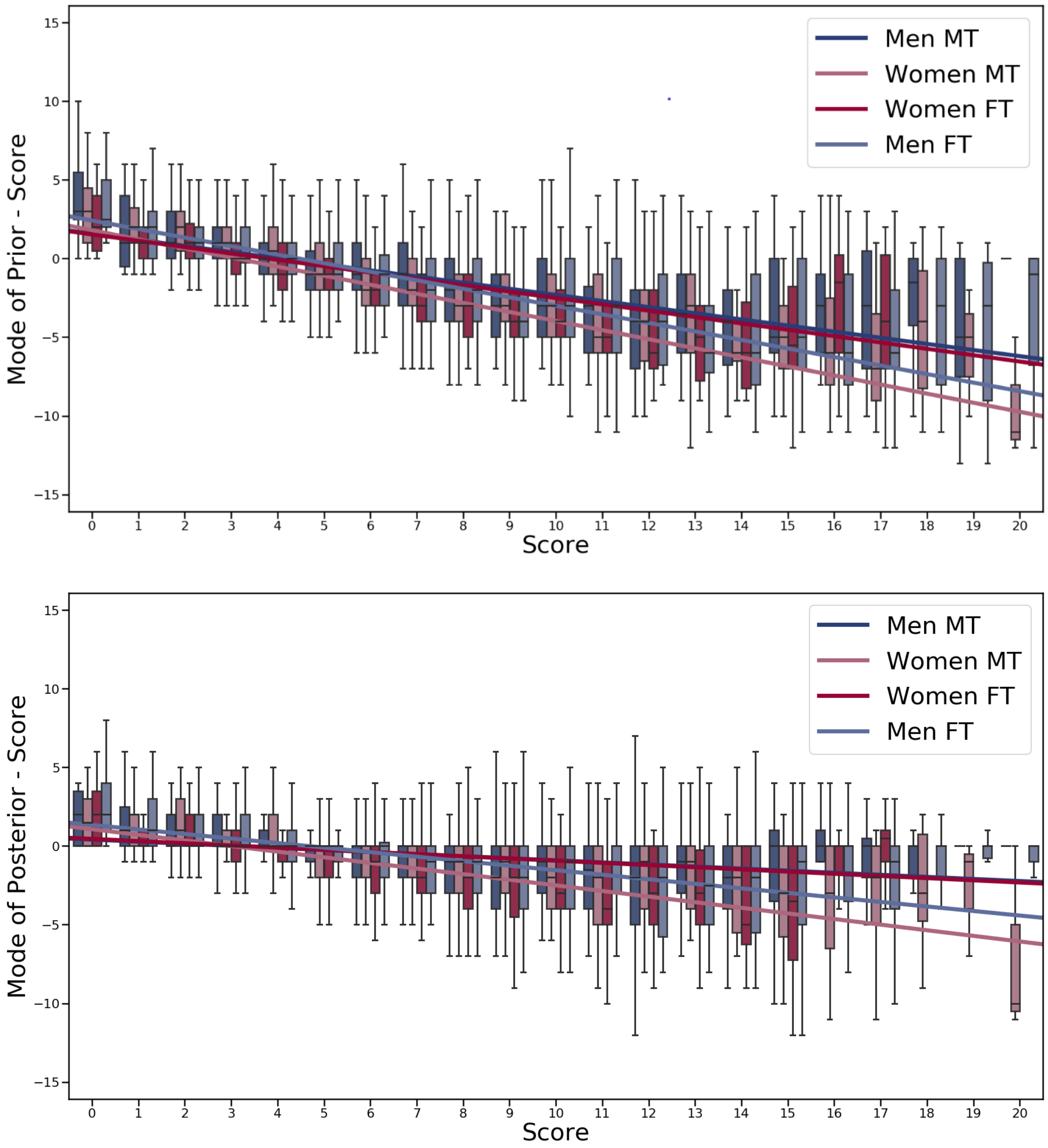

Figure B2. Boxplots of Prior and Posterior Beliefs in Study 2

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 lines represent linear fits of observed overconfidence on score.

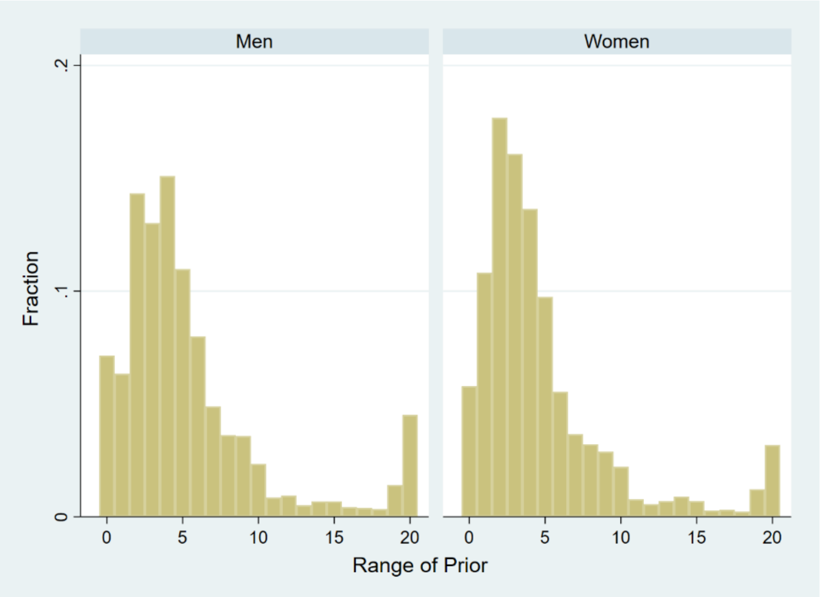

Figure B3. Range of Priors by Gender

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 A. 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	-1.84 (1.334)	-1.32 (1.051)	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 $t \neq t'$.[18] 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.[19] 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 $t \neq t^*$.[20]

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.[21] 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

[18] 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.

[19] In Study 1, we explore q = 0.6 and q=0.9. In Study 2, we explore q = 0.5 and q = 0.7.

[20] 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.

[21] 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:

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

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:

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

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.[22]

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.

[22] 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	0.38**** (0.032)	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 (2014), 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**** (0.029)	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.[23] 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%

[23] 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.