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

January 8, 2019

**Abstract:** We explore how beliefs respond to noisy information about own ability across a range of tasks, with a particular focus on how gender stereotypes impact belief updating. Participants in our experiments take tests of their ability across different domains. Absent feedback, beliefs of own ability are strongly influenced by gender stereotypes. We then provide noisy feedback about own absolute performance to participants and elicit posterior beliefs. Gender stereotypes have significant predictive power for posterior beliefs, both through their influence on prior beliefs (as predicted by a Bayesian model) but also through their influence on updating. Both men and women's beliefs are more responsive to information in gender congruent domains than gender incongruent domains. This is primarily driven by differential reactions to exogenously-received good news about own ability: both men and women react more to good news when it arrives in a gender congruent domain than when it arrives in a gender incongruent domain. Our results have important implications for understanding how feedback shapes gender gaps in self-assessments.

[‡] Corresponding author: kcoffman@hbs.edu.

Acknowledgements: Thank you to the NSF for their generous funding of this work.

## Introduction

Beliefs about own ability are a key input into many economically significant decisions. They shape financial decision-making (Barber and Odean 2001) and educational choices, such as what schools to apply to (Pan 2018) and what programs to study (Buser, Niederle, Oosterbeek 2014). They also likely impact labor market outcomes, shaping how job candidates present themselves (Reuben et al 2014) and what opportunities they apply to (Coffman, Collis, and Kulkarni 2018). Across each of these contexts, there exist varying degrees of uncertainty and ambiguity, creating the space for biases to flourish.

One important source of bias is gender stereotypes. 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 policy question is how might we reduce these gaps?[1] Perhaps the most obvious solution is providing increased information. If a student is unsure of her abilities in STEM, could her school 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, could her manager provide a more detailed performance review that addressed these areas? 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 a theoretical and empirical investigation of the effectiveness of increased information in reducing gender gaps in beliefs of own ability. In particular, we run a series of experiments that explore how individuals update their beliefs about themselves in response to noisy but quite informative

[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? While not the focus of this paper, this is another important issue to wrestle with.

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 2018). Is it also the case that these same gender stereotypes have predictive power for how individuals incorporate new information into their beliefs?

In our first simple 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 noisy information. 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. 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.

Across our two signal treatments, we find that, conditional on measured ability, women's priors of their own absolute test scores are approximately 0.58 points (or 0.13 standard deviations of ability) lower than men's, a statistically significant gap. Providing noisy but informative signals of own ability does not significantly reduce this gender gap. After the provision of signals, the gender gap remains at 0.52 points (or 0.12 standard deviations of ability). Interestingly, the gender gap directionally increases with feedback in the 90% Signal Accuracy treatment, suggesting that moving to more informative signals of own ability does not more effectively close the gap.

Our results motivate us to consider what underlies the ineffectiveness of information in closing the gender gap in beliefs. We design a second study to tease apart different mechanisms. In particular, we ask what the role is for gender stereotypes in predicting responses to information. We use a similar experimental paradigm, but expand the range of domains we consider. Rather than focus on cognitive ability, we test participants across eight different domains, chosen to vary in their associated gender stereotype.

Participants in the second study complete three rounds of an experiment very 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 explicitly collect data not only on their point-wise prior and posterior 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 speak to gender differences in the shape of belief distributions, and create a Bayesian benchmark for updating behavior for each participant. We build a simple model that offers predictions as to how a Bayesian would respond to the signal in our environment. This framework allows us to parse the results we observe, and uncover systematic departures from this 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, individuals' beliefs of their own ability increase significantly as the category becomes more gender congruent (i.e. more male-typed for men, or more female-typed for women). 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 our Bayesian model) but also through their impact on how participants react to information given the same prior. 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.

Because of our signal structure, we are able to explore how updating varies depending upon whether the signal draw was "good" or "bad" news relative to the truth. Importantly, in our design, the assignment of good or bad news through the randomly-drawn noise term is completely exogenous to both true ability and prior beliefs. We find that reactions to good and bad news vary with the gender stereotype of the domain. Both men and women are more responsive to good news when it arrives in gender congruent domains than when it arrives in gender incongruent domains. And, in gender incongruent domains, both men and women are more responsive to bad news than good. Our results suggest that convincing people of their talent in gender incongruent domains may be particularly 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. Most prior studies have focused on beliefs of relative ability, using coarse paradigms to allow for clean and practical tests of Bayesian models. 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 distributions over placement possibilities and compare updating behavior to a full Bayesian benchmark (see, for instance, Mobius et al 2014, Barron 2016, Buser et al 2016, 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). One main takeaway from this literature is that participants update differently about themselves than they do when forming beliefs about arbitrary statistical processes, though the differences are not entirely consistent across study. Some studies have found evidence for a "good news – bad news effect", with participants responding more to positive signals than negative signals (Eil and Rao 2011, Mobius et al 2014), though this finding is not universal.[2] Similarly, there is evidence that people are more overconfident 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, Gehards, and van der Weele 2016), consistent with theoretical models of motivated reasoning (such as Rabin and Schrag 1999, Benabou and Tirole 2002, or Koszegi 2006).

Results on gender within this literature have been mixed. Mobius et al (2014) report that women demonstrate more conservatism than men in their ego-relevant context, updating less in response to information, but that there are no gender differences in 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 what underlies the across-study differences. In recent work, Shastry, Shurchkov, and Xia (2018) explore how noisy feedback on relative ability shapes tournament entry decisions in a male-typed domain. They find that negative feedback is a strong deterrent of tournament entry for high ability women, primarily because they are too likely to attribute this feedback to ability rather than luck.

In our study, we will attempt to unpack the role of stereotypes in driving gender differences in beliefs. In addition, by focusing on absolute ability rather than relative ability, our experiment will yield rich data on good and bad news. Participants across the ability spectrum will be equally likely to receive exogenous 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.[3] Our results suggest that past findings of greater female conservatism could possibly be explained by (i) a sampling of primarily more male-typed domains, and (ii) an under-appreciation of gender differences in variance in priors. Our framework also helps provide additional insights into the mixed results on gender

[2] In particular, Ertac (2011) finds that participants respond more to bad news than good in her ego-relevant task.

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

differences in asymmetry. Our results suggest that whether or not a good news – bad news asymmetry is observed is mostly a function of how gender congruent the domain is, much less so 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, noisy information 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 the Bayesian model would predict. Rather, gender stereotypes seem to impact the way new information is incorporated into beliefs, fueling persistence in gender gaps.

## 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 or skipped questions are not penalized.

### *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 – $5^{th}$ percentile, $5^{th}$ – $20^{th}$ percentile, $20^{th}$ – $40^{th}$ percentile, $40^{th}$ – $60^{th}$ percentile, $60^{th}$ – $80^{th}$ percentile, $80^{th}$ – $95^{th}$ percentile, $95^{th}$ – $100^{th}$ 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.* 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 "Ɛ", where "Ɛ" 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 "Ɛ" the computer chose. We then give them a few examples of how different scores, draws of balls from the bag, and values of "Ɛ" 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. 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 (2018). 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 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. In Appendix Table B1, we present summary statistics on our workers by gender.

## 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 $p<0.001$.[4] 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 (4.16 SD) questions correctly on average, while women believe they answered 7.26 (3.86 SD) questions correctly on average ($p<0.001$). Recall that in this study participants are asked to guess what score they believe they earned. For simplicity, we'll refer to this as their "prior", despite the fact that it is a point prediction rather than a distribution.

In Table I, we regress these prior beliefs of both absolute and relative ability on participant gender and performance. Conditional on performance, we estimate that women state beliefs of absolute score approximately 0.6 points lower than men (Column I, $p<0.01$). Similarly, women believe they place 7.5 percentage points worse in the ability distribution compared to equally able men (Column II, $p<0.001$). Interestingly, even when we control for a participant's pointwise prior belief of her absolute score, women believe they place worse in the distribution than men do (Column III). 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.

[4] According to the model of stereotyping in Bordalo et al (2016) and as applied in their work on beliefs about gender (Bordalo et al (2018)), 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.

**Table I. Gender Differences in Prior Beliefs of Absolute and Relative Ability for Cognitive Skills Test**

| | OLS Predicting Prior Belief of Absolute Score | OLS Predicting Prior Belief of Percentile of Score | |
|---|---|---|---|
| | I | II | III |
| Female | -0.58***<br>(0.20) | -0.075****<br>(0.015) | -0.061****<br>(0.014) |
| Score | 0.60****<br>(0.022) | 0.023****<br>(0.0017) | 0.008****<br>(0.0021) |
| Prior Belief of Absolute Score | | | 0.024****<br>(0.002) |
| Demographic Controls | Yes | Yes | Yes |
| Constant | 2.77****<br>(0.70) | 0.33****<br>(0.053) | 0.27****<br>(0.050) |
| R-squared | 0.48 | 0.25 | 0.32 |
| N | 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 variance in Columns II and III.

***Posterior Beliefs***

Following the elicitation of priors, we provide a noisy signal of performance to participants. For participants in the 60% signal treatment, they receive a signal of their score that is their true score with probability 0.6. For participants in the 90% signal treatment, they receive a signal of their score that is their true score with probability 0.9. For those that do not receive a signal equal to their true score, they receive a signal that is their score plus an integer chosen from a uniform distribution over {-5,-4,-3,-2,-1,1,2,3,4,5}. We then elicit a posterior belief of absolute score from participants in the two signal treatments. Again, this is a point prediction, their belief of the score they most likely earned on the test.

There remains a significant gender gap in beliefs conditional on signal received across both signal treatments. 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).

In a Bayesian setting, these differences in posterior beliefs must be driven by differences in prior beliefs. A first step toward exploring the role of prior beliefs is to repeat this analysis while controlling for the priors that participants report. Pooling over the two signal treatments, we estimate that about half of the remaining gender gap is explained by 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 21% of the 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 prior beliefs of most likely score 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 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, true ability, and signal received – are predictive of posterior beliefs. We perform this analysis in Column III, including interactions of female with prior, true ability, 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.

**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) | -0.25** (0.12) | -0.16 (0.30) | -0.67*** (0.23) | -0.22 (0.17) | -0.42** (0.19) | -0.33** (0.16) |
| Score | Yes | Yes | 0.11** (0.054) | 0.42**** (0.056) | 0.15**** (0.045) | 0.38**** (0.082) | 0.18** (0.071) |
| Signal Received | Yes | Yes | 0.45**** (0.046) | 0.41**** (0.049) | 0.38**** (0.037) | 0.52**** (0.078) | 0.51**** (0.066) |
| Prior Belief of Absolute Score | | Yes | 0.47**** (0.030) | | 0.51**** (0.028) | | 0.36**** (0.026) |
| Female x Score | | | 0.068 (0.073) | | | | |
| Female x Signal | | | -0.17*** (0.064) | | | | |
| Female x Prior | | | 0.12*** (0.038) | | | | |
| Dummy for 90% Signal | Yes | Yes | Yes | | | | |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Treatment & Interactions | | | | | | | |
| 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. In Columns I – III, we interact the signal treatment dummy with score, signal received, and prior (in Columns II and III only) to allow for the fact that these variables may have differential impact across treatment.

***Whose Posterior Beliefs are More Accurate?***

We can also consider whether participants are updating in the direction of the truth – that is, are their posterior beliefs more accurate on average than their prior beliefs? In Figure I, we graph the average overconfidence for men and women by score, comparing posteriors and priors. In prior beliefs, average underconfidence increases with performance, consistent with work on overconfidence such as Moore and Healy (2008), with gender differences remaining relatively constant throughout the ability distribution.

We see that the signals are highly effective at moving individuals in the direction of the truth, reducing overconfidence among low ability individuals and reducing underconfidence among higher ability individuals. Indeed, 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 to 1.64 points ($p<0.001$). As expected, 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 true score, signal received, and prior belief, women's mean errors in posteriors are 0.35 points larger than men's ($p<0.01$). Relative to the accurate beliefs benchmark, men's and women's posteriors are still too underconfident on average, with participants updating in the correct direction but not enough.

The figure also makes it clear that the signals are not particular effective at closing the average gender gap in overconfidence, conditional on measured ability. Overall, we estimate that the gender gap in beliefs falls from 0.58 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 noisy but informative signals 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 stays relatively flat, at approximately 0.12 standard deviations of ability, in posteriors. Of course, the fact that signals are dramatically reducing the mean error in beliefs implies that the gender gap relative to mean error in beliefs is increasing rather substantially. 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 raise a number of important open 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? And, most centrally, to the extent that we see gender differences in prior distributions and updating, what underlies these differences? In particular, what is the role of gender stereotypes in shaping not only prior beliefs of ability, but also in predicting belief updating? To provide answers to these questions, we first develop a theoretical framework. Then, we report the results of 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.

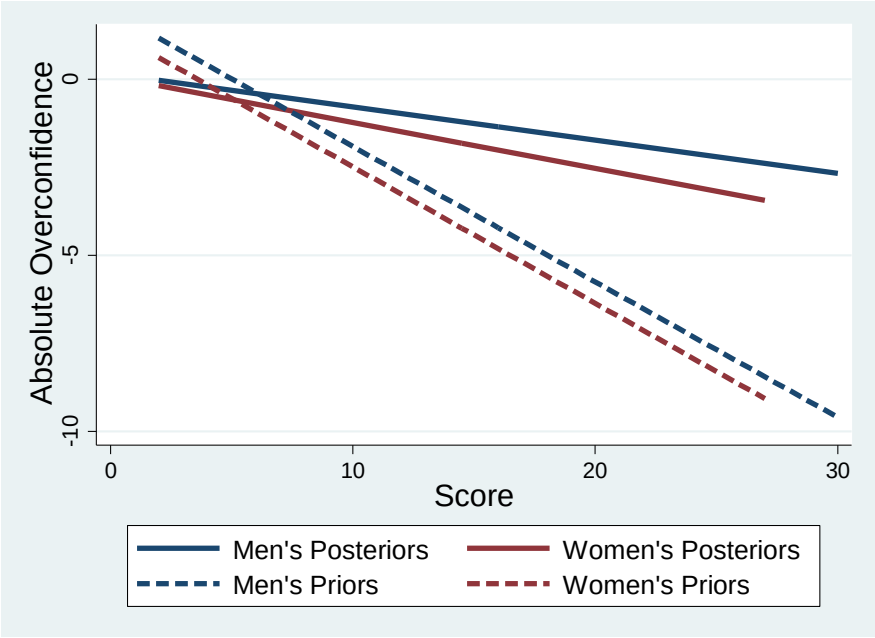


**Figure I. Adjustment in Beliefs**

**Theoretical Framework**

Denote the decision-maker's true score on the test by T, where T $\epsilon$ {0,1,2, ... , 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'.[5] The decision-maker then receives a signal of her performance, X, where X $\epsilon$ {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.[6] 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*.[7]

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 $\epsilon$ {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.[8] 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

[5] 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.
[6] In Study 1, we explore q = 0.6 and q=0.9. In Study 2, we explore q = 0.5 and q = 0.7.
[7] 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.
[8] 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 $\epsilon$ {-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?

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

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 $\epsilon$ {-5,-4,-3,-2,-1,1,2,3,4,5}, to produce:

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

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:

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

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:

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

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:

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

Plugging in,

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

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, we would have to that there existed 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:

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

Or, more simply:

$$r(T = t_j) > r(T = t')$$

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 ≠ j and $t_j$, $t_k$ in {Z-5, Z-4, Z-3, Z-2, Z-1, Z, Z+1, Z+2, Z+3, Z+4, Z+5}.

Plugging in,

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

Or, more simply:

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

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

*Proposition 2.* Suppose t* ≠ 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'.

With these results in mind, let's return to our experimental design. In the experimental framework of Study 1, we elicit the mode of the prior, t', and the mode of the posterior, t*. We also observe T, q and Z. Thus, for participants for whom t' in {Z-5, Z-4, Z-3, Z-2, Z-1, Z, Z+1, Z+2, Z+3, Z+4, Z+5}, the Bayesian model offers a sharp prediction. Note that this condition holds for 79% of our participants, including 78% of men and 80% of women. After seeing X=Z, these participants should report either their signal, Z, or the mode of their prior, t', as the mode of their posterior. For participants who see a large surprise - that is, t' > Z+5} or t' < Z+5} - the Bayesian model can justify reporting any value in {Z-5, Z-4, Z-3, Z-2, Z-1, Z, Z+1, Z+2, Z+3, Z+4, Z+5}. Note that, in this case, the Bayesian model cannot support reporting the mode of the prior as the mode of the posterior.

With this model in mind, we turn to Study 2. In this study, we will elicit complete information about participants' prior distributions over possible scores, allowing us to derive Bayesian benchmarks for behavior using the model above.

***Design of Study 2***

Study 2 builds on the paradigm of Study 1, but explores a variety of domains and collects additional data on prior and posterior beliefs. Building on the approach of Coffman (2014) and Bordalo et al (2018), 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. 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. 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 their believed most likely score, participants are asked on the next page how likely they think it is that they earned exactly this 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, we ask participants what their believed rank is, comparing themselves to 100 other randomly-chosen participants who completed the same multiple-choice test as part of the experiment.[9]

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.

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

[9] 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.

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 the experiment. They are given a slider scale that ranges from – 1 (women know much more) to 1 (men know much more) and 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 a target of 2,025 participants (25 of which participated one day ahead of the full HIT to ensure the functionality of the programming) in October 2018. We exclude four participants who failed a basic attention check (failing to identify a picture within a time limit).

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 round 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. 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 these specific instructions 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 suggests that that error did not impact the answers given to the questions. A full analysis of this issue is presented in Appendix C.

**Results for Study 2**

Appendix Table B2 presents summary statistics for our participants in Study 2. We control for all demographic characteristics in our regression going forward. We chose the categories for Study 2 to vary in their associated gender stereotype, as measured by actual and perceived gender gaps in ability. Table III confirms that our categories vary significantly along these dimensions. We arrange the categories by the

average slider scale rating given for the category among all participants. Four categories are perceived as being more female-typed: Kardashians, Disney, Art and Literature, and Verbal Skills, while four categories are perceived as being more male-typed: Business, Videogames, Sports, and Cars. The average gender gaps in performance correspond quite closely to these perceptions. In fact, the correlation between the male advantage in observed performance the observed average slider scale rating is 0.88.

**Table III. Average Gender Gaps and Slider Scale Perceptions**

| | **Kard-ashians** | **Disney** | **Art** | **Verbal** | **Business** | **Video-games** | **Sports** | **Cars** |
|---|---|---|---|---|---|---|---|---|
| Male Avg. Score | 8.27 | 8.16 | 7.35 | 4.50 | 5.64 | 10.93 | 7.93 | 8.39 |
| Female Avg. Score | 10.6 | 11.45 | 7.66 | 4.16 | 4.72 | 8.14 | 6.16 | 7.48 |
| Avg. Male Advantage | -2.34**** | -3.29**** | -0.31 | 0.34* | 0.93**** | 2.79**** | 1.77**** | 0.91**** |
| Avg. Slider Scale Rating | -0.55 | -0.38 | -0.18 | -0.18 | 0.19 | 0.44 | 0.48 | 0.52 |
| *N* | 765 | 757 | 759 | 749 | 754 | 744 | 768 | 767 |

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 of men and women.

***Prior Beliefs***

We start by exploring the prior beliefs of participants. Our goal is to understand what gender gaps look like, conditional on true performance, and what role stereotypes play in predicting these gender differences. A key question then is how to measure the impact of stereotypes. We follow the approach of Bordalo et al (2018). Under this model, a decision-maker's belief about herself is shaped in part by comparisons of the distribution of performance of her own gender in a category compared to the distribution of 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.

Practically, we can test for a role for this type of stereotypical thinking 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 ability. 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.

In Table IV, we explore gender differences and self-stereotyping in prior beliefs of absolute ability and relative ability. First, in Column 1, 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 row 3 of Table III), her observed performance, 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), round fixed effects, and category fixed effects. We estimate that, holding own ability fixed, women report prior beliefs approximately 0.5 points lower than men's for a gender-neutral category (average gender gap in performance of 0). More centrally, we see a strong role for stereotypes in shaping beliefs about self, replicating Bordalo et al (2018). We estimate that a 1-point increase in own gender advantage, roughly the size of the gender gap in business, decreases women's beliefs about their own ability by 0.16 points, while increasing men's beliefs about their own ability by 0.16 points. 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 that she reported, the results are nearly identical.[10]

**Table IV. Gender Differences in Priors of Absolute and Relative Ability**

| | **OLS Predicting Prior Belief of Score** | **OLS Predicting Prior Belief of Relative Rank (1=Best; 100=worst)** | |
|---|---|---|---|
| | I | II | III |
| Female | -0.49***<br>(0.099) | 4.86****<br>(0.86) | 3.74****<br>(0.84) |
| Own Gender Advantage | 0.16****<br>(0.022) | -0.78****<br>(0.18) | -0.40**<br>(0.18) |
| Score | 0.60****<br>(0.013) | -2.18****<br>(0.11) | -0.82****<br>(0.14) |
| Prior Belief of Absolute Score | | | -2.28****<br>(0.16) |
| Demographic Controls | Yes | Yes | Yes |
| R-squared | 0.45 | 0.13 | 0.19 |

[10] See Appendix Table B3 for these results. There, we also 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 0.72 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 0.14 points and increases men's beliefs of their own ability by 0.14 points.

| | | | |
|---|---|---|---|
| Clusters (Obs.) | 2021 (6063) | 2021 (6063) | 2021 (6063) |

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, fixed effects for category, and number of ASVAB questions answered correctly.

In Columns II and III we turn to beliefs of relative ability, predicting a participant's believed rank among 100 participants who completed the quiz in that category (1=best, 100=worst). We estimate that women believe they place nearly 5 ranks worse than men conditional on having the same score in a gender-neutral category. And, we see a significant role for own gender advantage. In Column III we add the participant's prior belief of her absolute score as a predictor. We replicate our finding from Study 1: a significant role for stereotypes remains after controlling for beliefs of own absolute ability, suggesting gender gaps in relative ability are also driven in part by biased beliefs of *others* ability.

In Figure II, we graph absolute overconfidence (believed score – observed score) against true score for both men and women. We split the sample by whether the observation is from a male-typed or female-typed domain. As in Study 1, we observe underconfidence on average in our sample, but with overconfidence for lower ability participants and growing underconfidence for higher ability participants. Perhaps the most striking result from Figure II 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. Rather, we see more underconfidence from women than men only in male-typed domains. When we focus on female-typed domains, women are actually less underconfident than men.

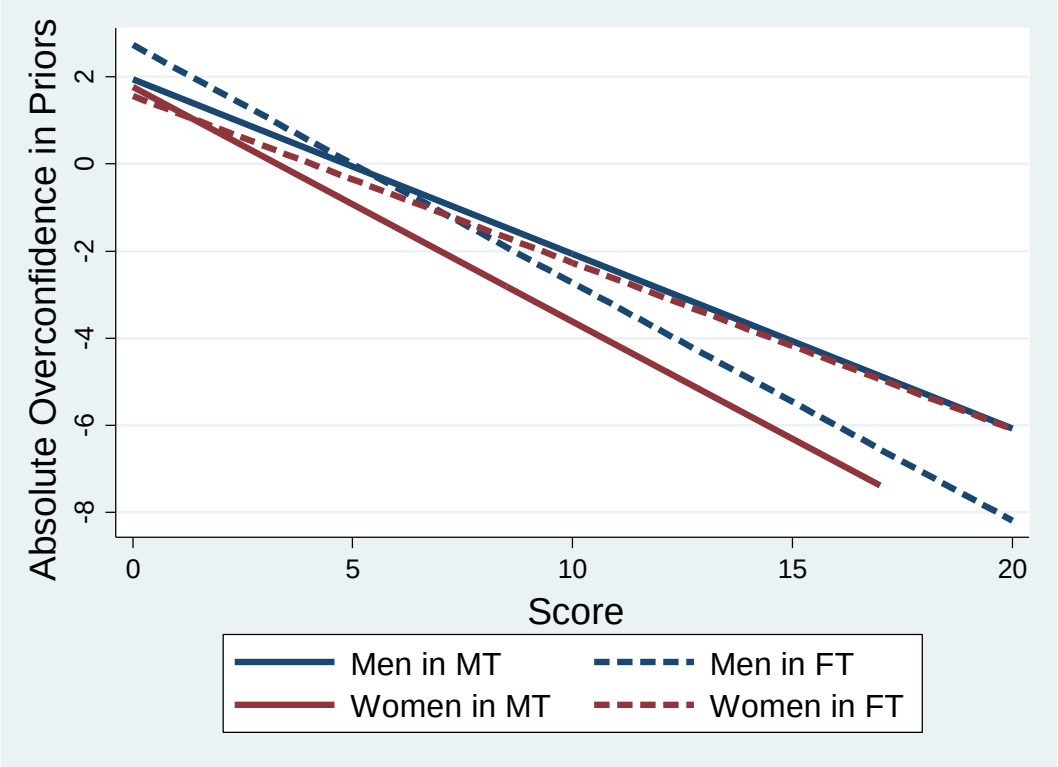


**Figure II. Average Overconfidence in Prior Beliefs**

***Exploring the Shape of Prior Distributions***

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 asymmetry (skewness).

Our first finding is that women have tighter, lower variance priors than men on average. In Figure B1 in the Appendix, we display a histogram of the ranges of priors in our sample, split by gender. We define the range 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 5.1. On average, women report narrower ranges than men (4.8 versus 5.5, $p<0.001$ clustering at individual level). We see a very similar pattern if we consider the individual standard deviations of priors. The average standard deviation of a prior distribution in our sample is 1.39, with women reporting lower variance distributions than men on average (average SD of 1.32 versus average SD of 1.51, $p<0.001$). Column I of Table V also indicates that gender stereotypes seem to play a small role in

predicting the variance in priors, with participants reporting more variant priors for more gender-congruent categories. This operates orthogonally to the main effect of female.

Next, we turn our attention to asymmetry or skewness in the prior. We start by defining three buckets of asymmetry in priors. We will 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. We see modest gender differences in asymmetry that do not appear to be a function of gender stereotypes. Women are approximately 3pp less likely to report a symmetric prior than a man (Column II, $p<0.05$), and approximately 2.5pp more likely to report a right-skewed prior (Column IV, $p<0.10$). For all measures of asymmetry, gender stereotypes are un-predictive.

Thus, while stereotypes are strongly predictive of the mode/mean of priors, they are not very predictive of their shape. We observe that women report tighter priors than men, with lower ranges and standard deviations on average. And, women are marginally more likely to report a right-skewed prior.

**Table V. Analysis of Gender Differences and Stereotypes in Predicting Shapes of Priors**

| | **OLS Predicting SD of Prior** | **OLS Predicting a Dummy if Prior Mean = Prior Median** | **OLS Predicting a Dummy if Left-skewed (Mean < Median)** | **OLS Predicting a Dummy if Right-skewed (Mean > Median)** | **OLS Predicting Right-Skewness of Prior (Mean – Median)** |
|---|---|---|---|---|---|
| | I | II | III | IV | V |
| Female | -0.19**** (0.052) | -0.03** (0.013) | 0.005 (0.015) | 0.025* (0.015) | 0.028* (0.016) |
| Own Gender Advantage | 0.016* (0.009) | 0.002 (0.003) | 0.001 (0.003) | -0.003 (0.003) | 0.000 (0.004) |
| Score | -0.02*** (0.006) | -0.004*** (0.002) | 0.004** (0.002) | 0.000 (0.002) | 0.000 (0.002) |
| Demographic Controls | Yes | Yes | Yes | Yes | Yes |
| R-squared | 0.06 | 0.02 | 0.01 | 0.02 | 0.02 |
| Clusters (Obs.) | 2021 (6063) | 2021 (6063) | 2021 (6063) | 2021 (6063) | 2021 (6063) |

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, fixed effects for category, and number of ASVAB questions answered correctly.

***Posterior Beliefs***

Do signals reduce the gender gap in beliefs, and in particular the reliance on stereotypes in shaping beliefs about own ability? Table VI replicates Table IV, but instead predicting posterior beliefs. In addition, we control for the signal received and signal treatment. Directionally, the estimated gender gap for a gender-neutral category and the reliance on stereotypes shrink in posterior beliefs of absolute ability compared to prior beliefs. However, both remain sizable and significant. While a 1-point increase in own gender advantage increased beliefs of own ability by an estimated 0.16 points in priors, the same shift in own gender advantage increases beliefs of own ability in posteriors by 0.14 points ($p<0.001$).[11] It is clear stereotypes continue to impact posterior beliefs.

We can also ask about "spillovers" onto beliefs of relative ability. Recall that our information provision focuses on absolute ability, providing a signal of absolute score on the test. However, one might expect that moving beliefs about absolute ability in the direction of the truth would also help to close the gender gap and the extent of stereotyping in beliefs of relative ability. In Column II we see that the signals do not reduce the gender gap in beliefs of relative ability, or the reliance on stereotypes in our data, as the coefficients on both are nearly identical in prior and posterior beliefs. Column III again asks about the role of beliefs of absolute score in explaining gender differences in beliefs of relative ability, and we continue to see that stereotypes are predictive even conditional on posterior believed score.

**Table VI. Posterior Beliefs of Absolute and Relative Ability**

| | **OLS Predicting Posterior Belief of Score** | **OLS Predicting Posterior Belief of Relative Rank (1 = Best, 100 = Worst)** | |
|---|---|---|---|
| | I | II | III |
| Female | -0.33***<br>(0.089) | 5.64****<br>(0.84) | 4.82****<br>(0.82) |
| Own Gender Advantage | 0.14****<br>(0.020) | -0.77****<br>(0.18) | -0.42**<br>(0.17) |
| Score | 0.54****<br>(0.026) | -1.80****<br>(0.23) | -0.47**<br>(0.24) |

[11] 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. We can also explore the shape of posteriors, paralleling our analysis of prior beliefs. We find that both the degree of variance and the degree of asymmetry in prior beliefs are strongly predictive of variance and asymmetry, respectively, in posteriors. Conditional on the shape of priors, we see no gender differences or differences by gender stereotypes in shape of posteriors. Results available upon request.

| | | | |
|---|---|---|---|
| Posterior Belief of Absolute Score | | | -2.48**** (0.17) |
| Demographic Controls | Yes | Yes | Yes |
| R-squared | 0.64 | 0.17 | 0.23 |
| Clusters (Obs.) | 2021 (6063) | 2021 (6063) | 2021 (6063) |

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, fixed effects for category, 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.

Figure III replicates Figure II, this time plotting average overconfidence in posteriors (differencing the mode of the posterior and true score). We can clearly see that while average overconfidence has shrunk, relative to overconfidence in priors, the gender gaps are quite similar in size when compared to prior beliefs.

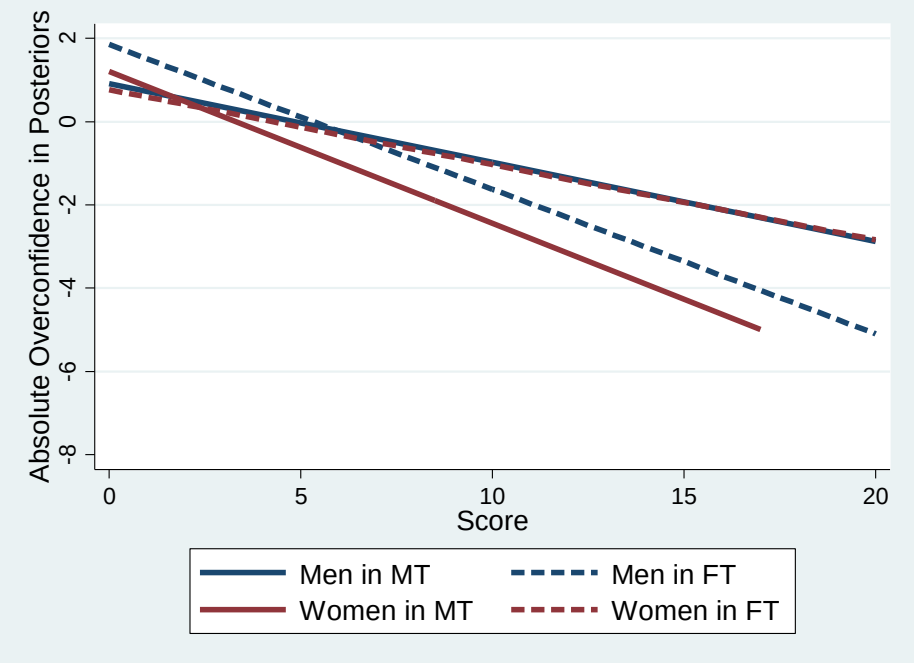


**Figure III. Overconfidence in Posterior Beliefs**

***How Well Does the Bayesian Model Predict Posterior Beliefs?***

One important question is whether these remaining gender differences are explained by gender differences in prior beliefs. 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 role for priors. We can use the theoretical framework of Section III and the full distribution of prior beliefs over possible scores provided by each participant to

generate a "Bayesian prediction" of the mode of the posterior distribution that should be reported by each participant. This Bayesian prediction tells us, given the signal structure, the prior beliefs reported by the participant, including the probability mass she assigns to each possible score, and the signal she receives, what score would a Bayesian decision-maker report as the mode of their posterior. We do this for each person-round observation in our dataset.

For a small fraction of our participants, the Bayesian prediction is unclear: these are the participants who put no positive 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.

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 or the mode of the 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 (see Proposition 2).[12] Note that these proportions do not sum to 100% because for some participants, the signal is also the mode of their prior.

With these Bayesian predictions, we can create a counterfactual – what would the population of posterior beliefs look like if every participant updated in accordance with Bayes rule? In a Bayesian world, would signals be sufficient to close the gender gap and reduce reliance on stereotypes? In Table VII, Column I below, we predict the Bayesian-predicted posterior for each observation from our standard explanatory variables – gender, own gender advantage, performance, demographics, and round and category fixed effects. We see that the Bayesian model predicts a modest gender gap in gender-neutral categories, and a significant effect of self-stereotyping. Relative to the impact of stereotyping seen in priors (an estimated coefficient of 0.16), Bayesian posteriors would have stereotyping of roughly 5/8 the magnitude (an estimated coefficient of 0.10). Given that the impact of stereotypes in our observed posteriors in Column I of Table VII is larger than this Bayesian prediction (0.14, or 7/8 the magnitude of the effect in priors), we

[12] 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 (176 of the 579 observations in this category), we take the average of those modes as the Bayesian prediction.

have reason to suspect that there are departures from the Bayesian model in our data. Put differently, the effect of stereotypes is approximately 40% larger in our observed data than what the Bayesian model would predict.

We can explore departures from the Bayesian predictions by predicting an individual's reported posterior belief from her Bayesian prediction. Column II presents the results. While the Bayesian prediction does have significant predictive power for observed posteriors, we see systematic departures from Bayes that are well-predicted by gender stereotypes. That is, we estimate a significant role for 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. For every 1 point increase in own gender advantage, we estimate that beliefs of own performance, *conditional on the Bayesian prediction*, increase by 0.10 points. That is, even after fully accounting for how stereotypes would impact posteriors through the Bayesian impact of prior beliefs, we estimate that stereotypes have significant predictive power for posterior beliefs. Stereotypes seem to color the way participants update in response to noisy information.

Finally, in Column III, we ask whether the extent to which we observe departures from the Bayesian prediction and reliance on stereotyping varies by signal accuracy. We interact Bayesian prediction, female, and own gender advantage with a dummy for the 70% signal accuracy treatment. We see no significant differences by signal accuracy treatment.

**Table VII. Bayesian Predictions versus Observed Posteriors**

| | **OLS Predicting Bayesian-Prediction of Posterior** | **OLS Predicting Observed Posterior** | |
|---|---|---|---|
| | I | II | III |
| Female | -0.13*<br>(0.067) | -0.29***<br>(0.083) | -0.16<br>(0.108) |
| Own Gender Advantage | 0.10****<br>(0.018) | 0.10****<br>(0.019) | 0.12****<br>(0.027) |
| Score | 0.89****<br>(0.009) | 0.40****<br>(0.017) | 0.39****<br>(0.017) |
| Bayesian Prediction | | 0.44****<br>(0.017) | 0.43****<br>(0.019) |

| | | | |
|---|---|---|---|
| Signal 70 | | | 0.20<br>(0.19) |
| Signal 70 x Female | | | -0.25<br>(0.16) |
| Signal 70 x Own Gender Advantage | | | -0.038<br>(0.037) |
| Signal 70 x Bayesian Prediction | | | 0.031<br>(0.020) |
| Demographic Controls | Yes | Yes | Yes |
| R-squared | 0.72 | 0.68 | 0.68 |
| Clusters (Obs.) | 2021 (6063) | 2021 (6063) | 2021 (6063) |

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, fixed effects for category, and number of ASVAB questions answered correctly.

Figure IV helps us to better understand what predicts departures from Bayesian predictions in our setting. We graph the observed posterior belief on the y-axis against the Bayesian predicted posterior on the x-axis, graphing the fitted lines for both men and women in male-typed and female-typed domains. We also provide the 45-degree line as a point of reference. For both men and women, posteriors are more responsive to the Bayesian prediction when the domain is gender congruent. As we've seen in past figures, men and women in gender congruent categories behave quite similarly, with nearly identical average posterior beliefs conditional on the Bayesian prediction. In gender incongruent categories, both men and women hold relatively lower beliefs than what Bayes would predict compared to gender congruent categories, and their beliefs are less responsive to the Bayesian predictions.

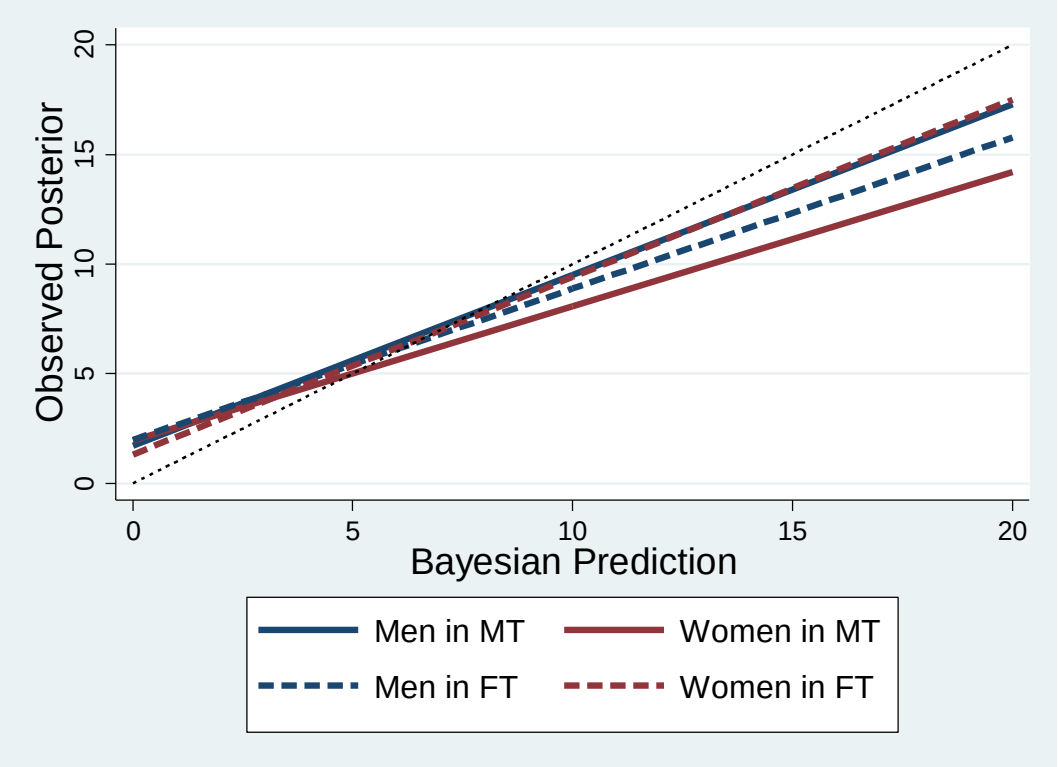


**Figure IV. Observed Posteriors Compared to Bayesian Predictions**

We formalize this finding in Table VIII, predicting posterior beliefs from the Bayesian predictions and splitting the sample by male-typed and female-typed categories. We estimate that women hold significantly lower beliefs than men conditional on the Bayesian prediction in male-typed categories (Column I); this is driven by the fact that 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.

**Table VIII. 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 | -0.75**** (0.10) | -0.072 (0.19) | 0.052 (0.11) | -0.65*** (0.22) | -0.36** (0.16) |

| | | | | | |
|---|---|---|---|---|---|
| Bayesian Prediction | 0.44**** (0.022) | 0.49**** (0.026) | 0.45**** (0.025) | 0.38**** (0.034) | 0.43**** (0.021) |
| Female x Bayesian Prediction | | -0.10**** (0.029) | | 0.10*** (0.031) | 0.001 (0.021) |
| Own Gender Advantage | | | | | -0.073** (0.034) |
| Own Gender Advantage x Bayesian Prediction | | | | | 0.023**** (0.004) |
| Score | 0.37**** (0.022) | 0.36**** (0.022) | 0.41**** (0.025) | 0.41**** (0.025) | 0.39**** (0.017) |
| Demographic Controls | Yes | Yes | Yes | Yes | Yes |
| R-squared | 0.64 | 0.64 | 0.71 | 0.71 | 0.68 |
| Clusters (Obs.) | 1860 (3033) | 1860 (3033) | 1870 (3033) | 1870 (3033) | 2021 (6063) |

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, fixed effects for category, and number of ASVAB questions answered correctly.

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 lead to different participant behavior in terms of posteriors. Consider an 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 see behaviorally 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 weight assigned to it, the weight assigned to the signal, the standard deviation of the prior, and the degree of positive skewedness – 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 Table IX, replicating our main results from the previous two tables (Column II of Table VII, Column V of Table VIII). We see that contrary to the theoretical prediction, 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 (higher SD) and are more positively-skewed on average report greater posteriors conditional on other observables. The hypothesis above, that some priors make it "easier" to follow the Bayesian prediction, seems to be supported by the data. Interestingly, the particular weight assigned to the mode of the prior or the signal ex-ante does not add predictive power beyond what the Bayesian prediction incorporates. It's simply the mode of the prior, likely proxying for how far away the Bayesian prediction is from the prior belief, and the shape of the prior, likely reflecting how uncertain the participant was, that seems to matter.

Conditioning on this additional information eliminates the estimated gender gap in posterior beliefs for a gender-neutral category; however, a significant, albeit smaller, role for stereotypes persists. It continues to be the case that, conditional on both the Bayesian prediction and more detailed information about participant priors, that own gender advantage shapes posterior beliefs. Column I indicates that own gender advantage leads to greater posterior beliefs of own performance, holding fixed all other factors, while Column II shows that this effect primarily operates through the fact that participants are more responsive to the Bayesian prediction as own gender advantage in the category increases.

**Table IX. Using Other Characteristics of the Prior to Predict Posteriors**

| **OLS Predicting Observed Posterior** | | |
|---|---|---|
| | I | II |
| Female | -0.03<br>(0.056) | -0.08<br>(0.10) |
| Bayesian Prediction | 0.03<br>(0.019) | 0.02<br>(0.022) |
| Female x Bayesian Prediction | | 0.002<br>(0.015) |

| | | |
|---|---|---|
| Own Gender Advantage | 0.04***<br>(0.014) | -0.06**<br>(0.025) |
| Own Gender Advantage x Bayesian Prediction | | 0.013****<br>(0.003) |
| Score | 0.40****<br>(0.015) | 0.39****<br>(0.015) |
| Mode of Prior | 0.61****<br>(0.021) | 0.61****<br>(0.021) |
| Weight on Mode of Prior | 0.00<br>(0.00) | 0.00<br>(0.00) |
| Weight on Signal in Prior | 0.00<br>(0.00) | 0.00<br>(0.00) |
| SD of Prior | 0.09***<br>(0.032) | 0.09***<br>(0.031) |
| Right-skewedness of Prior | 0.28****<br>(0.065) | 0.28****<br>(0.065) |
| Noise Draw | 0.24****<br>(0.016) | 0.24****<br>(0.016) |
| Demographic Controls | Yes | Yes |
| R-squared | 0.81 | 0.81 |
| Clusters (Obs.) | 2021<br>(6063) | 2021<br>(6063) |

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, fixed effects for category, and number of ASVAB questions answered correctly.

### *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) in ego-relevant tasks. With our data, we can ask, in an absolute ability framework, whether a similar asymmetry is observed. Furthermore, 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.[13] Because the signal displayed is exogenous conditional on performance, this definition of news avoids any selection on priors or performance. Of course, these

[13] 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).

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.

We start with a simple graphical exploration of the data. In Figure V below, we plot the difference between a participant's observed posterior and the Bayesian prediction for her posterior against her true score. Positive values indicate a participant reports a posterior belief that exceeds the Bayesian prediction of her posterior. We split the data into two panels, graphing Male-Typed Domains in Panel (a) and Female-Typed Domains in Panel (b). Within each panel, we have four types of observations: men and women who receive good and bad news.

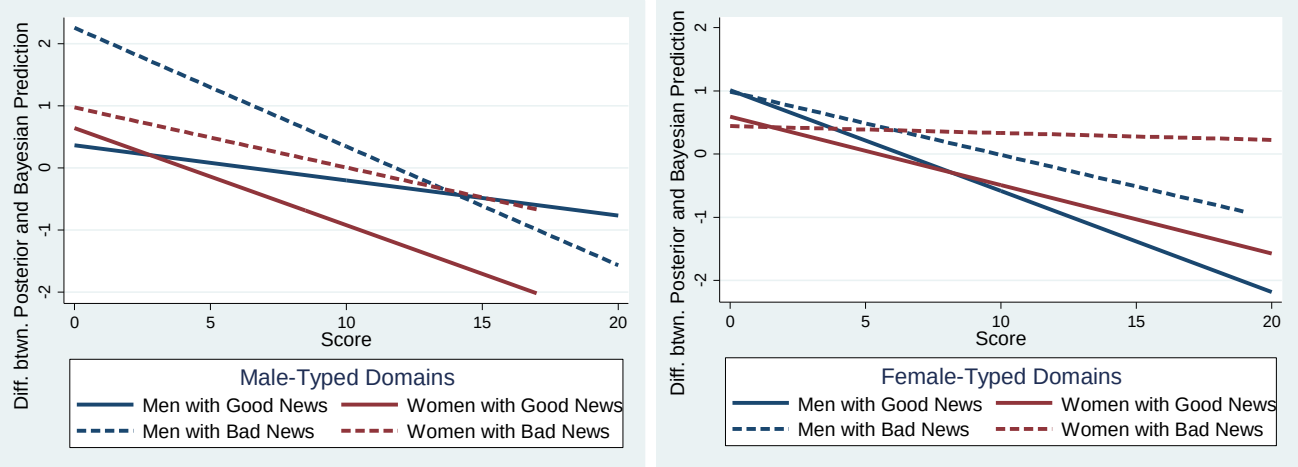


**Figure V. Good News and Bad News Reactions**

The first observation is that departures from the Bayesian prediction depend upon the type of news received. For both men and women in both types of domains, observed posteriors after receiving bad news on average exceed the Bayesian prediction by more than observed posteriors after receiving good news. Consistent with previous work that has found evidence of conservatism (Eil and Rao 2011, Mobius et al 2014), we on average see "under-shooting" of the Bayesian prediction: those who receive good news on average report posteriors that are too low relative to the Bayesian prediction, while those who receive bad news on average report posteriors that are too high relative to the Bayesian prediction.

There also appear to be differences by gender stereotype. In the male-typed domains, we see that men's beliefs after receiving good news on average exceed women's beliefs after receiving the same type of news (relative to the Bayesian predictions). This is not true in female-typed domains, where men and women's beliefs after receiving good news look more similar. A similar pattern emerges for bad news. Men's beliefs after receiving bad news in a male-typed domain are, for a large range of scores, on average greater than women's beliefs after receiving bad news. In the female-typed domains, this pattern is again reversed. This

evidence is suggestive that who reacts *more* to good or bad news is a function of the gender stereotype of the domain.

We follow the econometric approach of Eil and Rao (2011) in exploring these patterns formally. 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 posterior would predict for this particular type of news relative to when the same Bayesian prediction is made for bad news. 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.

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 (which is purely exogenous), 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 (mirroring Figure V). The results are presented in Table X.

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 0.75). However, men have a larger estimated intercept for bad news than women do (a constant of 2.21 as opposed to 1.43). This suggests that after receiving bad news, men report larger 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 0.80 for men and 0.59 for women). It seems to be the case that women discount increasingly good news more than men do when it comes in a male-typed domain. The estimated constant for women receiving good news is larger than the estimated constant for men (2.07 compared to 1.54), suggesting that for those whose Bayesian prediction is quite close to 0 women will be directionally more overconfident than men after receiving good news.

**Table X. Good News and Bad News**

| | OLS Predicting Posterior Belief | | | |
|---|---|---|---|---|
| | Male-Typed Domains | | Female-Typed Domains | |
| | Men | Women | Men | Women |

| | | | | |
|---|---|---|---|---|
| Bayesian Prediction | 0.75**** (0.051) | 0.75**** (0.040) | 0.78**** (0.053) | 0.92**** (0.022) |
| Good News Dummy | -0.67 (0.43) | 0.64** (0.25) | 0.75* (0.43) | 0.50** (0.24) |
| Good News x Bayesian Prediction | 0.045 (0.056) | -0.16**** (0.048) | -0.11* (0.063) | -0.13**** (0.028) |
| Constant | 2.21**** (0.40) | 1.43**** (0.21) | 1.39**** (0.35) | 0.90**** (0.21) |
| R-squared | 0.63 | 0.46 | 0.52 | 0.72 |
| Cluster (Obs.) | 753 (1240) | 1107 (1793) | 735 (1172) | 1135 (1858) |
| | | | | |
| Estimated Responsiveness to Good News | 0.80 | 0.59 | 0.67 | 0.79 |
| Estimated Responsiveness to Bad News | 0.75 | 0.75 | 0.78 | 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.

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 0.78). We also estimate a smaller constant for women than men after receiving bad news (0.90 versus 1.39). 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 0.79 and 0.67 for women and men, respectively). 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 prediction increases.

To summarize, gender stereotypes play an important role in predicting asymmetry. Men and women are similarly responsive to good news when it comes in gender congruent domains. But, both men and women are much less responsive to that same good news when it instead comes in a gender incongruent domain.

In fact, for individuals operating in gender incongruent domains, responsiveness to bad news consistently exceeds responsiveness to good news.

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 (0.63 for men in male-typed domains, 0.72 for women in female-typed domains). But, when individuals are operating in gender incongruent domains, posteriors are much harder to predict (r-squared of 0.46 for women in female-typed domains, r-squared of 0.52 for men in female-typed domains).

In Appendix Table B5, 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 those participants who are in some sense most over and under-confident in their prior beliefs. 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, unlike in the full sample where we estimate greater responsiveness to bad news than good in gender incongruent domains, we estimate similar responsiveness to good and bad news in gender incongruent domains for this sub-sample, suggesting some of the asymmetry in gender incongruent domains may be driven by people who receive "large" surprises.

In Appendix Table B6, 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.

**Conclusion**

There is increasing evidence that stereotyped beliefs drive important economic decisions – such as willingness to answer when unsure (Baldiga 2014, Coffman and Klinowski 2018), willingness to contribute ideas (Coffman 2014, Chen and Houser 2017, Bordalo et al 2018), willingness to compete (Niederle and Vesterlund 2007), and willingness to lead (Born et al 2018). Given this growing consensus, an important question is how beliefs evolve over time. In this paper, we take an important step toward addressing this

question, asking how individuals respond to noisy but highly informative feedback about their abilities across different domains.

We replicate previous findings in showing 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 in a way that is well-predicted by 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. Within the context of our cognitive skills setting, more informative feedback is no more effective (and is in fact directionally less effective) in closing the gender gap.

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 much 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. 2016. “Belief Updating: Does the ‘good-News, Bad-News’ Asymmetry Extend to Purely Financial Domains?” Working Paper SP II 2016-309. WZB Discussion Paper. https://www.econstor.eu/handle/10419/147020.

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.

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 Seff-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 B. Coffman, Nicola Gennaioli, and Andrei Shleifer. 2018. “Beliefs about Gender.” *American Economic Review* forthcoming.

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. 2018. “A Man’s World? The Impact of a Male Dominated Environment on Female Leadership.” *Working Paper*, November. http://papers.ssrn.com/abstract=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.

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. 2018. “Gender and Promotion.” *Working Paper*.

Coffman, Katherine B., and David Klinowski. 2018. “The Impact of Penalties for Wrong Answers on the Gender Gap in Test Scores.” *Working Paper*, September.

Coutts, Alexander. 2018. “Good News and Bad News Are Still News: Experimental Evidence on Belief Updating.” *Experimental Economics* forthcoming.

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.

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.

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.

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

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.

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. 2018. “Deception and Self-Deception.” *Working Paper*, April.

Shastry, G. Kartini, Olga Shurchkov, and Lingjun “Lotus” Xia. 2018. “Luck or Skill: How Women and Men Respond to Noisy Feedback.” *Working Paper*.

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.

Siqi, Pan. 2017. “The Instability of Matching with Overcon Dent Agents: Laboratory and Field Investigations.” *Games and Economic Behavior* forthcoming (November).

## 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 | 0.09 |
| Black | 0.08 | 0.09 | 0.54 |
| Asian | 0.08 | 0.05 | <0.01 |
| Attended HS in US | 0.96 | 0.97 | 0.71 |
| HS Only | 0.10 | 0.09 | 0.51 |
| Some College/Assoc. | 0.33 | 0.39 | <0.01 |
| Bachelors | 0.42 | 0.38 | 0.08 |
| Advanced Degree | 0.16 | 0.14 | 0.31 |
| ASVAB Score (out of 5)[ǂ] | 3.39 | 3.39 | 0.88 |
| Treatment Assignment | | | |
| 50% Signal | 0.48 | 0.51 | 0.18 |
| 70% Signal | 0.52 | 0.49 | 0.18 |
| *N* | 804 | 1217 | |

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

**Table B3. Robustness Analysis of Prior Beliefs**

| | OLS Predicting Prior Belief of Score – Mode of Prior | OLS Predicting Prior Belief of Score – Mean of Prior | OLS Predicting Prior Belief of Score – Mode of Prior | OLS Predicting Prior Belief of Score – Mean of Prior |
|---|---|---|---|---|
| | I | II | III | IV |
| Female | -0.49*** (0.099) | -0.45**** (0.105) | -0.49**** (0.099) | -0.45**** (0.105) |
| Own Gender Advantage in Performance | 0.16**** (0.022) | 0.18**** (0.022) | | |
| Own Gender Advantage in Perception | | | 0.72**** (0.095) | 0.76**** (0.099) |
| Score | 0.60**** (0.013) | 0.61**** (0.014) | 0.60**** (0.013) | 0.61**** (0.014) |
| Demographic Controls | Yes | Yes | Yes | Yes |
| R-squared | 0.45 | 0.43 | 0.45 | 0.43 |
| Clusters (Obs.) | 2021 (6063) | 2021 (6063) | 2021 (6063) | 2021 (6063) |

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, fixed effects for category, and number of ASVAB questions answered correctly.


**Figure B1. Range of Priors by Gender**

**Table B4. Robustness Analysis of Posterior Beliefs**

| | **OLS Predicting Posterior Belief of Score – Mode of Posterior** | **OLS Predicting Posterior Belief of Score – Mean of Posterior** | **OLS Predicting Posterior Belief of Score – Mode of Posterior** | **OLS Predicting Posterior Belief of Score – Mean of Posterior** |
|---|---|---|---|---|
| | I | II | III | IV |
| Female | -0.33*** (0.089) | -0.42**** (0.097) | -0.33**** (0.089) | -0.41**** (0.097) |
| Own Gender Advantage in Performance | 0.14**** (0.020) | 0.14**** (0.020) | | |
| Own Gender Advantage in Perception | | | 0.66**** (0.085) | 0.63**** (0.089) |
| Score | 0.54**** (0.026) | 0.54**** (0.027) | 0.54**** (0.026) | 0.54**** (0.027) |
| Demographic Controls | Yes | Yes | Yes | Yes |
| R-squared | 0.64 | 0.60 | 0.64 | 0.60 |
| Clusters (Obs.) | 2021 (6063) | 2021 (6063) | 2021 (6063) | 2021 (6063) |

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, fixed effects for category, 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.

**Table B5. Good News and Bad News for Only Those Who Had Clear Bayesian Prediction**

| | OLS Predicting Posterior Belief | | | |
|---|---|---|---|---|
| | Male-Typed Domains | | Female-Typed Domains | |
| | Men | Women | Men | Women |
| Bayesian Prediction | 0.80**** (0.051) | 0.82**** (0.028) | 0.81**** (0.052) | 0.97**** (0.019) |
| Good News Dummy | -0.89** (0.44) | 0.17 (0.19) | 0.34 (0.45) | 0.47** (0.21) |
| Good News x Bayesian Prediction | 0.14*** (0.054) | 0.004 (0.035) | -0.002 (0.063) | -0.053** (0.023) |
| Constant | 1.74**** (0.42) | 0.99**** (0.15) | 1.19**** (0.36) | 0.41**** (0.18) |
| R-squared | 0.74 | 0.67 | 0.60 | 0.83 |
| Cluster (Obs.) | 717 (1130) | 1047 (1627) | 707 (1087) | 1071 (1661) |
| | | | | |
| Estimated Responsiveness to Good News | 0.94 | 0.82 | 0.81 | 0.92 |

| | | | | |
|---|---|---|---|---|
| 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.

**Table B6. 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 | 0.78**** (0.034) | 0.82**** (0.031) | 0.71**** (0.059) | 0.89**** (0.021) |
| Good News Dummy | -0.73** (0.36) | 1.07**** (0.23) | -0.30 (0.49) | 0.28 (0.23) |
| Good News x Bayesian Prediction | 0.01 (0.042) | -0.27**** (0.041) | -0.02 (0.067) | -0.11**** (0.028) |
| Constant | 2.14**** (0.32) | 1.19**** (0.19) | 2.12**** (0.45) | 1.09**** (0.20) |
| R-squared | 0.64 | 0.48 | 0.52 | 0.72 |
| Cluster (Obs.) | 753 (1240) | 1107 (1793) | 735 (1172) | 1135 (1858) |
| | | | | |
| Estimated Responsiveness to Good News | 0.79 | 0.55 | 0.69 | 0.78 |
| Estimated Responsiveness to Bad News | 0.78 | 0.82 | 0.71 | 0.89 |

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.

**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 | OLS Predicting Probability Assigned to Mode of Prior | OLS Predicting Probability Assigned to Mode of Posterior |
|---|---|---|---|
| Instruction Error was Fixed | -1.58<br>(1.31) | -1.22<br>(1.04) | 0.11<br>(1.01) |
| Demographic Controls | Yes | Yes | Yes |
| R-squared | 0.02 | 0.03 | 0.01 |
| Cluster (Obs.) | 2021 (2021) | 2021 (6063) | 2021 (6063) |

Notes: * indicates significance at p<0.10, ** at p<0.05, *** at p<0.01, and **** at p<0.001.