The Gender Gap in Confidence: Expected But Not Accounted For*

Christine L. Exley† Kirby Nielsen‡

Draft: October 31, 2022

Abstract

We investigate how the gender gap in confidence affects the views that evaluators (e.g., employers) hold about men and women. If evaluators fail to account for the confidence gap, it may cause overly pessimistic views about women. Alternatively, if evaluators expect and account for the confidence gap, such a detrimental impact may be avoided. We find robust evidence for the former: even when the confidence gap is expected, evaluators fail to account for it. This "contagious" nature of the gap persists across many interventions and types of evaluators. Only a targeted intervention that facilitates Bayesian updating proves (somewhat) effective.

KEYWORDS: gender, confidence, beliefs, experiments

*We thank Zoë Cullen, Oliver Hauser, Judd Kessler, Muriel Niederle, Ryan Oprea, Emma Ronzetti, Colin Sullivan, Lise Vesterlund, and many seminar participants for helpful comments and suggestions. This study was reviewed by the Institutional Review Board at Harvard Business School.

†clexley@hbs.edu; Harvard Business School

‡kirby@caltech.edu; California Institute of Technology

1 Introduction

Women are underrepresented and underpaid in many areas of the labor market, especially in male-stereotyped fields (Bertrand and Katz, 2010; Goldin, 2014; Blau and Kahn, 2017; Michelmore and Sassler, 2016). A large body of work has identified factors that may contribute to these gender gaps. Review articles highlight gender differences in the willingness to negotiate (Hernandez-Arenaz and Iriberri, 2019) and compete (Niederle and Vesterlund, 2011; Niederle, 2016), gender differences in risk preferences (Croson and Gneezy, 2009), and the role of discrimination (Riach and Rich, 2002). Recent papers further narrow in on factors such as female leaders being rewarded less than equally-effective male leaders (Grossman et al., 2019), women requesting lower starting salaries than men (Roussille, 2021), women being less likely to self-report qualifications (Murciano-Goroff, 2021), and women negotiating less even in a female-dominated profession (Biasi and Sarsons, Forthcoming).

One of the literature's most robust findings is the gender gap in confidence (Lundeberg et al., 1994; Mobius et al., 2022), even among elite academics (Sarsons and Guo, 2021) and especially in male-stereotyped fields (Beyer, 1990; Bordalo et al., 2019; Coffman et al., 2019a; Exley and Kessler, 2022). Many papers highlight how the confidence gap may affect the "supply" of women in the labor market. For example, the confidence gap relates to women having lower earnings expectations (Reuben and Zafar, 2017), being less likely to enter competitive fields (Niederle and Vesterlund, 2007; Buser et al., 2014), being less likely to speak up (Coffman, 2014), and being less likely to apply for challenging work (Coffman et al., 2019c). But, less is known about how the confidence gap affects the "demand" for women, which is the focus of this paper.

How the confidence gap may affect the demand of women is unclear. On one hand, if others expect the confidence gap—perhaps due to movements such as "Lean-in"—then they may account for it in a way that ensures that women's relative underconfidence does *not* cause overly pessimistic beliefs about women. On the other hand, if others—such as employers, colleagues, and peers—do not expect or do not account for the confidence gap when forming beliefs about men and women, then the confidence gap will be "contagious." For instance, the confidence gap may cause others to form overly pessimistic beliefs about women when reviewing job applications in which the candidate discusses their own performance and ability, when making promotion decisions that are in part based off of self-evaluations, and when selecting leaders and team members based off their self-reported qualifications. More pessimistic beliefs about women may, in turn, contribute to worse outcomes for women and may exacerbate gender discrimination (Bohren et al., 2019; Coffman et al., 2021).[1]

[1]There are also many other important factors, e.g., the relative weight placed on luck (Erkal et al., 2021).

To first establish that there is a confidence gap in our setting, "workers" complete a math and science test and then answer 17 self-evaluation questions about their performance on the test. Workers are incentivized to accurately answer each self-evaluation question. The confidence gap proves robust: across the 17 self-evaluation questions—and significantly so in 16 of these questions— female workers provide more pessimistic beliefs about their performance than equally performing male workers do. For instance, when focusing on our main sample of workers for which there is no actual gender difference in performance, answers to our main self-evaluation question reveal that 80% of women believe they have a "poor performance" while only 56% of men do. Workers know that they are classified as having a "poor performance" if another randomly-selected participant who does not know their gender deems the number of questions they got right on the test as being "poor."

Then, to investigate how this confidence gap affects others' beliefs about men and women, we investigate whether "evaluators" account for the confidence gap after learning how workers answer the main self-evaluation question. Specifically, after evaluators learn whether they will be asked to provide incentivized beliefs about a randomly-selected male worker or instead a randomly-selected female worker (who we refer to as "their worker"), the *Baseline* treatment involves five main stages.[2] First, we elicit evaluators' *prior* by asking them to guess the percent chance that their worker has a poor performance. Second, we provide evaluators with *accurate* aggregate information about workers' self-evaluations: evaluators who are asked to provide beliefs about a randomly-selected female worker are informed that 80% of female workers thought they had a poor performance, and evaluators who are asked to provide beliefs about a randomly-selected male worker are informed that 56% of male workers thought they had a poor performance. Third, to examine how this information influences evaluators' beliefs, we elicit their *posterior* about the percent chance that their worker has a poor performance. Fourth, to investigate whether the confidence gap is expected, we elicit evaluators' beliefs about their worker's *overconfidence* and *underconfidence* by asking them to guess the percent chance that their worker is overconfident conditional on having a poor performance and the percent chance that their worker is underconfident conditional on having a good performance. Finally, evaluators answer additional incentivized questions that measure their susceptibility to cognitive biases and complete a short follow-up survey.

Before receiving any information on workers' self-evaluations, evaluators expect that female workers are slightly more likely than male workers to have a poor performance. According to their priors, however, this expected performance gap is small (∼3.9 percentage points) and is not statistically different from the true gap (∼1.7 percentage points).

[2] As explained in Section 2.2, we ask about a subgroup of workers for whom there are no actual gender differences in performance. But, as shown in Sections 6.8 and 6.9, our results are not reliant on this restriction.

After evaluators receive information on workers' self-evaluations—information that conveys more pessimistic views held by female workers or more optimistic views held by male workers—does this expected performance gap become substantial because evaluators fail to account for the confidence gap in these self-evaluations? That is, does the confidence gap prove to be "contagious"? Or, is the potentially detrimental impact of the confidence gap avoided because evaluators expect and account for the confidence gap?

Two results *seem* to point towards the latter at first blush. First, as indicated via their beliefs about workers' confidence, evaluators expect the confidence gap in self-evaluations. Evaluators expect that, among workers with a poor performance, male workers are 8.25 percentage points significantly more likely than female workers to be overconfident and incorrectly guess that they have a good performance. Evaluators also expect that, among workers with a good performance, female workers are 10.07 percentage points significantly more likely than male workers to be underconfident and incorrectly guess that they have a poor performance. Second, we can calculate—from evaluators' priors, the information on workers' self-evaluations, and evaluators' beliefs about the accuracy of that information (given their beliefs about workers' confidence)—the posterior beliefs that evaluators would hold if they were Bayesian. These *implied Bayesian posterior beliefs* indicate that the confidence gap should not be contagious, and specifically, that the information on workers' self-evaluations should not result in overly pessimistic views about women.

We nonetheless find the opposite to be true. The confidence gap in workers' self-evaluations is contagious. After receiving information on workers' self-evaluations, evaluators hold overly pessimistic view about the relative performance of women. According to their posteriors, evaluators now expect a large and statistically significant performance gap ($\sim$10.5 percentage points). This expected performance gap is indeed 6 times larger than the true performance gap and nearly 3 times larger than evaluators' priors. Thus, the confidence gap exacerbates the expected performance gap, even though Bayesian updating implies that it shouldn't and even though the confidence gap is expected when asked about directly.

We explore the robustness of this result across many dimensions. Our results persist when evaluators make predictions about: other types of performance outcomes (e.g., absolute and relative performance rather than subjective performance), workers who face strategic incentives to inflate their self-evaluations, and specific workers after learning their specific self-evaluation. Our results also persist when we manipulate the salience of a worker's gender, when more information is known about worker quality, when evaluators gain more experience with self-evaluations, and when restricting to the subset of evaluators who expect the confidence gap. In addition, when we elicit beliefs about male and female workers simultaneously, we find that our results persist on the individual-level: while only 39% of evaluators expect

a performance gap according to their priors, 79% of evaluators expect a performance gap according to their posteriors. Finally, and related to recent work on understanding experts' beliefs (DellaVigna and Pope, 2018a,b), we replicate our results with professional evaluators who self-report hiring and managerial experience. Relative to our main evaluators, professional evaluators differ slightly in their priors because they accurately expect no—rather than a small—performance gap. Nonetheless, according to their posteriors, the confidence gap in self-evaluations causes these professional evaluators to inaccurately expect just as large of a performance gap relative to our main evaluators.

While it is clear that evaluators do not *accurately* update in response to the information on workers' self-evaluations—and hence do not accurately adjust for the confidence gap in self-evaluations—the question remains as to whether evaluators update in a manner that accounts, at least to some degree, for whether they are provided with self-evaluations from men or women. To investigate this, in the *Unknown Gender* treatment, evaluators are asked about either "group-1" or "group-2" workers. Evaluators know that workers are assigned to these groups based on their answers to a question in a follow-up survey, but they do not know what this question is (and in particular, do not know that the question is about the worker's gender). Strikingly, evaluators' posteriors indicate an expected performance gap that is statistically indistinguishable from the expected performance gap when evaluators instead know the gender of workers. Thus, we have no evidence for evaluators accounting for *gender differences* in confidence when forming their posteriors.

In considering what prevents evaluators from accounting for the confidence gap when forming their posterior beliefs, our results have already shown that it is *not* because the confidence gap is unexpected. That said, it could be that evaluators are not attending to the confidence gap when forming their posterior beliefs. To investigate this possibility, we test a light-touch intervention. In our *Attention* treatment, we make beliefs about confidence more salient by eliciting evaluators' confidence beliefs before—rather than after—their posterior beliefs. This intervention proves ineffective: the expected performance gap remains at the same (substantial and significant) level.

Another possibility is that evaluators—despite expecting the confidence gap—are either unable or unwilling to exert the effort needed to adjust for the confidence gap when forming their posterior beliefs. To investigate this possibility, we test a much more extensive intervention. In our *Calculation* treatment, to alleviate any difficulty with Bayesian updating, we provide evaluators with their implied Bayesian posterior beliefs before eliciting their actual posteriors. This intervention proves effective: the expected performance gap shrinks and is only marginally significantly different from the true performance gap. Intriguingly, and consistent with our results operating through a cognitive channel, the effectiveness of this

intervention aligns with us also observing that the extent to which evaluators' posteriors disfavor women is positively and significantly correlated with exhibiting base rate neglect.

In summary, we document that evaluators *expect* a confidence gap, but they do not account for it. The confidence gap conveyed via self-evaluations causes evaluators to (inaccurately) form overly pessimistic beliefs about the performance of women than men. This "contagious" nature of the confidence gap persists across a variety of conditions, interventions, and types of individuals—with only a targeted intervention that helps evaluators with Bayesian updating proving (somewhat) effective.

To better understand the potential impact of gender differences in the labor market, our work complements the aforementioned rich literature on how the confidence gap affects the decisions made by men and women *themselves* by additionally examining how the confidence gap affects *others' beliefs* about men and women. Our work is thus related to the small but growing body of literature on how the confidence gap affects others' decisions—and hence may relate to others' beliefs—about men and women. This literature shows that the confidence gap conveyed via group interactions may relate to women being selected less frequently as leaders (Reuben and Zingales, 2012), that the confidence gap conveyed via workers' self-reported beliefs may explain why providing these self-reports to employers does not mitigate their male hiring preference (Reuben et al., 2014), and that the confidence gap conveyed with employees' self-evaluations does not influence employers' relative ratings of their male and female employees (Bohnet et al., 2022).

Relative to this literature, part of our main contribution lies in eliciting a variety of incentivized beliefs that allow us to cleanly document that and narrow in on *why* individuals do not account for confidence gap.[3] Indeed, our evidence makes clear that neither expecting the confidence gap nor having greater attention drawn to the confidence gap nor believing that one accurately adjusted for the confidence gap is sufficient for evaluators to account for it. Our evidence further shows that the confidence gap results in overly pessimistic beliefs about women relative to men even though—if evaluators were Bayesians—it should not. Relatedly, our evidence points towards the need for more extensive interventions that directly help individuals to adjust for the confidence gap and opens up many lines for future work that we return to in Section 7. In addition, by introducing conditions in which the gender of workers is unknown—a type of control that is often infeasible with field work—our paper is uniquely situated to narrow in on the extent to which evaluators adjust for the *gender* gap in confidence.[4]

[3] Reuben et al. (2014) have unincentivized belief data consistent with their findings. Reuben and Zingales (2012) and Bohnet et al. (2022) do not have belief data.

[4] A worker's self-evaluation may not influence others' beliefs if, e.g., others believe the self-evaluations are uninformative. To isolate reasons like these, the gender unknown condition is particularly useful.

2 Experimental Design

Our experimental design involves two main types of participants: (i) "workers" who complete a math and science test and provide self-evaluations about their performance on that test, and (ii) "evaluators" who make predictions about the performance of workers, sometimes after learning about the workers' self-evaluations.

Our main *Worker Study* involves one treatment described in Section 2.1. Our main *Evaluator Study* involves six treatments. While we describe the *Baseline* treatment of the *Evaluator Study* in Section 2.2, we describe the additional five treatments in Section 2.4. In between, to help motivate these additional treatments, we describe our main outcomes in Section 2.3. Finally, in Section 2.5, we describe recruitment and implementation details for these studies and provide an overview on the additional study versions. In total, we recruited 6,894 participants—mostly on Prolific and as detailed later.

2.1 Design for The *Worker Study*

The *Baseline* treatment of the *Worker Study* involves two main parts: Part 1 and Part 2. In addition to a $3 completion fee for a 15-minute study, workers may earn up to $1 in bonus payment, randomly selected from either Part 1 or Part 2.

In Part 1, workers answer a 10-question math and science test.[5] Workers have 15 seconds to answer each question, and workers are never provided with any information on their performance on this test. If Part 1 is selected as the part-that-counts, workers earn 10 cents for each question they answer correctly.

After Part 1 but before Part 2, workers select an answer that is equal to 0, 1, ..., or 10 questions in response to the following (unincentivized) "classifier question."[6]

- (CLASSIFIER QUESTION) An individual's performance on the math and science test was indicative of poor math and science skills if the number of questions the individual answered correctly was less than or equal to ____ .

In Part 2, workers answer 17 self-evaluations—displayed in random order—about their own performance. If Part 2 is randomly selected as the part-that-counts, then workers receive the amount they earn in one randomly selected self-evaluation and are incentivized to answer accurately.[7] We focus here on 1 of these 17 self-evaluations, called our *main self-evaluation*

[5] We selected ten questions from the Armed Services Vocational Aptitude Battery (ASVAB), which is used to assess aptitude in various technical fields. We tell participants that "performance on this test is often used as a measure of cognitive ability by academic researchers."

[6] Workers answered two classifier questions, but we focus here on the one that we use in the *Evaluator Study*. The full text of both questions can be found in Appendix Table A.1.

[7] See the table note of Appendix Table A.1 for details on randomization and incentives.

question, as this question is most relevant for understanding our main *Evaluator Study*. We describe the details of the other self-evaluation questions in Appendix Table A.1.

The *main self-evaluation question* is as follows:

- (Main Self-Evaluation Question) Did your classifier describe your performance as poor?

In response to the main self-evaluation question, workers can select "yes" or "no" and know that they earn $1 in that self-evaluation if their guess is correct. To answer the main self-evaluation question, workers are told that they will be matched with another worker (called their "classifier") who is equally likely to be a male worker or female worker.[8] We tell workers that their score is classified as "poor performance" if it was less than or equal to the threshold score that their classifier indicated in the Classifier Question described above. For example, if a worker's classifier says that an individual's performance is indicative of poor math and science skills if they answered 5 or fewer questions right, then the worker is classified as having a poor performance if they scored 0–5 on the test. While we will use this shorthand of "poor performance" throughout the rest of our paper for conciseness, we instead write out the definition of poor performance ("performance on the math and science test that was indicative of poor math and science skills") in the text of the questions provided to workers, as shown in Appendix Table A.1.[9]

2.2 Design for the *Baseline* treatment of the *Evaluator Study*

In the *Baseline* treatment of the *Evaluator Study*, each evaluator is incentivized to accurately answer four predictions about the performance of "their worker." Evaluators know that their worker will be randomly selected from the available pool of female workers (and thus referred to as "your female worker") or instead will be randomly selected from the available pool of male workers (and thus referred to as "your male worker"). Thus, evaluators are only asked about female workers *or* male workers. Each of the four predictions relates to whether their worker has a poor performance, defined in the same manner as noted in Section 2.1.[10]

In their first prediction, evaluators provide their *prior belief* about the percent chance that their worker has a poor performance.

[8] In the study, we actually refer to "classifiers" as "evaluators." But, to avoid confusion with our later study versions, we refer to them as classifiers in our paper.

[9] Specifically, the main self-evaluation question corresponds to Self-Evaluation 8B in Appendix Table A.1. In addition to the definition of poor performance being written out, note that the "classifier" is referred to as their "evaluator" as previously explained in Footnote 8.

[10] In the question text provided to evaluators, the definition of poor performance is written, and the worker's "classifier" is referred to as the worker's "evaluator" (see Appendix Table A.2).

- (Prior Belief) What do you think is the percent chance that your male/female worker in this prediction had a poor performance?

After their first prediction but before their second prediction, we provide evaluators with *accurate* information on how workers—in the available pool of workers from which their worker could be randomly selected—answered the main self-evaluation question. Specifically, evaluators are informed that 80% of female workers thought they had a poor performance if their worker is a randomly-selected female worker or instead that 56% of male workers thought they had a poor performance if their worker is a randomly-selected male worker. Then, we ask evaluators to provide their *posterior belief* about the percent chance that their worker has a poor performance.

- (Posterior Belief) After completing the math and science test, 56%/80% of male/female workers predicted that they had a poor performance. What do you think is the percent chance that your male/female worker in this prediction had a poor performance?

Finally, evaluators make their last set of predictions via a strategy-method style elicitation in which they are asked the following two predictions to assess how likely they think it is that their worker is overconfident or underconfident.

- (Overconfidence Belief) If your male/female worker in this prediction had a poor performance, what do you think is the percent chance that he/she is overconfident because he/she predicted that he/she did NOT have a poor performance?

- (Underconfidence Belief) If your male/female worker in this prediction did not have a poor performance, what do you think is the percent chance that he/she is underconfident because he/she predicted that he/she had a poor performance?

After evaluators make their predictions, we ask them five incentivized "bonus" questions designed to test for common cognitive biases that might correlate with belief updating behavior. We defer explanation of these questions to our discussion of the related results in Section 5.3. Then, evaluators complete a short follow-up survey to gather additional control and demographic information.

We conclude the main experimental design with two additional notes: one on the available pool of workers and another on incentives. On the available pool of workers, recall that evaluators make predictions about their male *or* female worker who is randomly selected from the available pool of workers. Evaluators are informed that this available pool of workers is the group of male/female workers who had performances in the "middle," or in

the 25th–75th percentile, in the *Worker Study*.[11] This restricted worker pool allows us to ensure that there are no gender differences in the actual performance of workers, which proves useful for analyses. That said, later study versions show that similar results persist when we do not have to rely on this restriction to ensure there are no gender differences in the actual performance—i.e., see the *Evaluator (Professional Evaluators) Study* in Section 6.1, the *Evaluator (Additional Demographics) Study* in Section 6.8, and the *Evaluator (Known Performance) Study* in Section 6.9.

On incentives, evaluators know they are equally likely to receive how much they earn from (i) their prior belief, (ii) their posterior belief, or (iii) either their overconfidence or underconfidence belief, depending on which of these two beliefs is relevant given the strategy-method elicitation. Evaluators report each belief in the form of a percent chance of some outcome being true (0-100%) and may earn a $1 bonus according to an incentive-compatible Becker-DeGroot-Marschak (BDM) procedure.[12] In addition, after making these predictions, evaluators are surprised with the opportunity to earn $1 if they correctly answer one randomly-selected bonus question.

2.3 Main Outcomes

Recall that we define a worker as having a "poor performance" if their score was classified as indicative of poor math and science skills in response to the Classifier Question. Related, we define a worker as having a "good performance" if their score was *not* classified as indicative of poor math and science skills in response to this question.

From the *Worker Study*, our main outcome relates to whether a worker predicts that they have a poor performance.

From the *Evaluator Study*, we have four main outcomes that we directly elicit from evaluators: (i) their *prior belief* about the chance that their (male or female) worker has a poor performance, (ii) their *posterior belief* about the chance that their worker has a poor performance—after they learn that 56%/80% of male/female workers predicted that they have a poor performance, (iii) their *overconfidence belief* about the chance that their worker predicts they have a good performance when they actually had a poor performance, and (iv) their *underconfidence belief* about the chance that their worker predicts they have a bad

[11] We describe this to evaluators as follows: "Workers who had performances in middle neither performed the best nor performed the worst. According to the number of questions they got right on the math and science test, workers who had performances in the middle performed better than or equal to at least one-quarter of all workers, and they performed worse than or equal to at least one-quarter of all workers."

[12] Specifically, they are told that to secure the largest chance of earning $1 from each self-evaluation, they should report their most-accurate guess. They are then allowed to click on a button to reveal the precise payment rule. For the 19% of participants who choose to reveal this information, they are provided with full details of the BDM procedure. For more on the BDM procedure, see Mobius et al. (2022).

performance when they actually had a good performance

The evaluators' main outcomes further imply an additional piece of information about evaluators' beliefs. As detailed in Appendix E, we can calculate an evaluator's *implied Bayesian posterior belief* given an evaluator's *prior belief*, *overconfidence belief*, and *underconfidence belief*. That is, after we provide evaluators with a signal of their worker's performance (i.e., that 56%/80% of male/female workers guessed that they had a poor performance), we can calculate what their posterior belief would be if they were Bayesian. This is because evaluators' overconfidence and underconfidence beliefs determine their beliefs about the accuracy of this self-evaluation information and hence how much they should update their prior beliefs.

The *implied Bayesian posterior beliefs* are of particular interest to us because, by comparing their *posterior belief* to their *implied Bayesian posterior belief*, we can examine the extent to which evaluators adjust for the gender gap in self-evaluations relative to the extent to which they should adjust for the gender gap in self-evaluations if they are Bayesian. In addition, we can also benchmark their beliefs in relation to the truth.

Finally, since we are interested in how workers' gender influences evaluators' beliefs, we will analyze the "gender gap" in these beliefs. Specifically, given that evaluators are randomly assigned to evaluate male *or* female workers, we analyze an across-evaluator measure of the difference in average beliefs about female workers compared to average beliefs about male workers (although, as shown in Section 6.5, our results are robust to a study version that allows us to elicit within-subject measures of beliefs about men and women).

2.4 Design for the other treatments of the *Evaluator Study*

In the *Evaluator Study*, evaluators are randomly assigned to either (i) make predictions about male workers only, or (ii) make predictions about female workers only. In addition, evaluators are randomized into one of the following six treatments: (i) the *Baseline* treatment, (ii) the *Attention* treatment, (iii) the *Calculation* treatment, (iv) the *Baseline, Unknown Gender* treatment, (v) the *Attention, Unknown Gender* treatment, and (vi) the *Calculation, Unknown Gender* treatment. While Section 2.2 detailed the *Baseline* treatment, this section will describe the remaining five treatments. See also Appendix Figures A.5 –A.7 for an overview of the timelines in each of these treatments.

To investigate a light-touch intervention intended to increase evaluators' attentiveness about their worker's confidence, the *Attention* treatment changes the order in which beliefs are elicited. In the *Baseline* treatment, we elicit evaluators' beliefs in the following order: (i) their prior beliefs, (ii) their posterior beliefs, and then (iii) their overconfidence and underconfidence beliefs. In the *Attention* treatment, the only difference is that the order changes

to be the following: (i) their prior beliefs, (iii) their overconfidence and underconfidence beliefs, and then (iii) their posterior beliefs.

We run the *Calculation* treatment to investigate a more extreme intervention designed to directly help evaluators update in response to information on the workers' self-evaluation. The only difference between the *Attention* and *Calculation* treatments is that, on the screen where evaluators report their posterior beliefs in the *Calculation* treatment, we provide them with their implied Bayesian posterior. We tell evaluators that this implied Bayesian posterior incorporates the information contained in the workers' self-evaluations together with the evaluators' own prior belief and over/underconfidence beliefs. See Appendix Figures A.3 and A.4 for an example of how the screen eliciting posterior beliefs changes in the *Calculation* treatment relative to the *Baseline* and *Attention* treatment.

Finally, to examine the role of knowing workers' gender in forming and updating beliefs—and hence to investigate whether evaluators account for *gender* differences in self-evaluations—we ran three additional treatments in which the gender of workers is not known. Specifically, for $X \in \{\text{Baseline, Attention, Calculation}\}$, the *X, Unknown Gender* treatment is the same as the X treatment except for the fact that the gender of the workers is unknown. In the *Unknown Gender* treatments, instead of making predictions about male or female workers, evaluators make predictions about "group-1" or "group-2" workers. We tell evaluators that a worker is assigned to group-1 or group-2 based on how they answered a follow up question, but we do not tell evaluators what this follow-up question is. In practice, we use the gender question from the follow-up survey, so group-1 workers are the exact same set as our male workers and group-2 workers are the exact same set as our female workers.

2.5 Implementation and Recruitment Details

In all of our studies, participants receive ample instructions and are required to correctly answer understanding questions before proceeding to the main parts of our study. Rather than excluding participants, they are given as many times as needed to correctly answer the understanding questions. For full experimental instructions of all study versions that we run, see the supplemental Online Appendix.

In April 2022, we recruited 403 participants on Prolific to complete our study as "workers."[13] After excluding 10 participants who neither identify as men nor women because we are under-powered to consider this group, this resulted in 393 workers in the *Baseline* treatment

[13]To be eligible for our study, participants needed to have completed at least 100 prior submissions on Prolific with an approval rating of 95% or greater and chose the United States as their residence. Also, since we recruited a gender balanced sample, participants must have selected either Male or Female for their sex on the Prolific platform—although we use their self-identified gender from our follow-up survey.

of the *Worker Study*.

In May 2022, we recruited 2,400 participants on Prolific to complete studies as "evaluators" (see footnote 13 for eligibility criteria). These evaluators were randomized into one of six treatments: the *Baseline* treatment (n=402), the *Attention* treatment (n=403), the *Calculation* treatment (n=405), the *Unknown Gender* treatment (n=405), the *Attention, Unknown Gender* treatment (n=392), and the *Calculation, Unknown Gender* treatment (n=393).

We recruited an additional 1,091 workers and 3,000 evaluators to complete additional study versions. We will discuss these study versions as the paper progresses, and we provide an overview in Appendix Table A.5.[14]

3 Worker Results

To establish the confidence gap, we first examine data from the *Baseline* treatment of the *Worker Study*, which we often refer to as simply the *Worker Study*. Out of the 10 questions on the math and science test, men slightly outperform women: on average, male workers answer 4.62 correctly and female workers answer 4.27 correctly ($p = 0.08$). However, when we restrict to the available pool of workers that evaluators are asked about (i.e., workers with performances in the "middle") men and women perform equally well: on average, male workers answers 4.49 correctly and female workers answer 4.41 correctly ($p = 0.58$).

As expected, similar results follow about likelihood of male and female workers having "poor performances," which is particularly important for the results of our *Evaluator Study* since evaluators are asked to make predictions about the chance of male and female workers having poor performance. When we consider all men and women, women are marginally significantly more likely to have a poor performance: the likelihood of a poor performance is 53% among female workers but only 47% among male workers ($p = 0.09$). But, when we restrict to the available pool of workers that evaluators are asked about, women and men are equally likely to have a poor performance: the likelihood of a poor performance is 49.53% among female workers and 47.79% among male workers ($p = 0.56$).[15]

Nevertheless, when restricting to this available pool of workers for which there is no performance gap, we find that female workers' self-evaluations indicate lower confidence than those of male workers. Table 1 presents results on how male and female workers answer

[14]Related to one of our additional worker and evaluator studies, we also recruited 100 participants to complete the study as "employers," as detailed in footnote 33.

[15]To calculate a worker's true chance of a poor performance, we determine the percent of classifiers who classified the worker's score as indicative of poor math and science skills in response to the Classifier Question. Then, to determine the chance that a randomly-selected male/female worker has a poor performance, we average these chances across all male/female workers.

the *main self-evaluation question* by showing the likelihood that a worker predicts that they have a poor performance (the dependent variable equals 1 if a worker predicts that they have a poor performance and 0 otherwise) regressed on *Female*, which is an indicator for female workers. Columns 1 and 2 present the results when considering all workers, while Columns 3 and 4 restrict to the available pool of workers. In addition, Columns 2 and 4 include performance fixed effects to allow us to assess gender differences in self-evaluations among equally performing men and women. In all cases, it follows that women are significantly more likely to predict that they have a poor performance. From this data, we can also see the information on workers' self-evaluations that we provide to evaluators. The estimates in Column 3 show that—among the available pool of workers—56% of male workers believe they have a poor performance (see the coefficient estimate on the constant) while 80% of female workers believe they have a poor performance (note the sum of the coefficient estimates on the constant and *Female*). We summarize these results below in Result 0.

Result 0 (The Confidence Gap)**.** *Female workers report significantly more pessimistic self-evaluations than equally performing male workers.*

Table 1: Self-Evaluations in the *Baseline* treatment of the *Worker Study*

	All Workers		Available Pool of Workers	
	(1)	(2)	(3)	(4)
Female	0.185***	0.155***	0.233***	0.232***
	(0.047)	(0.044)	(0.057)	(0.056)
Constant	0.573***		0.563***	
	(0.035)		(0.044)	
N	393	393	249	249
Perf FE	No	Yes	No	Yes

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are robust and shown in parentheses. Results are from OLS regressions of the responses provided to the main self-evaluation question, coded as 1 if the workers guess they have a "poor performance" and 0 otherwise. *Female* is an indicator for the worker identifying as a woman. Perf FEs are dummies for each possible performance out of the 10 questions on the test. In Columns 1–2, data are from the 393 participants who identified as a man or a woman in the *Baseline* Treatment of the *Worker Study*. In Columns 3–4, data are further restricted to the available pool of workers that evaluators are asked about—i.e., male and female workers with performances in the "middle" or 25th-75th percentile.

While the above focuses on documenting the confidence gap in response to the main self-evaluation question since our main evaluators are provided with information on how workers answer that question, recall that workers answered 16 other self-evaluation questions as well. Appendix Table B.1 presents the regression results of all self-evaluations. These results reveal that women report more pessimistic self-evaluations than equally-performing men in

all 17 self-evaluation questions, and significantly so in 16 of the 17. Women are less likely to believe they answered at least 3, 5, or 7 questions right; are less likely to believe they scored in the top half relative to men, women, or all other participants; and, are more likely to believe they have a poor performance according to two subjective classifications.

4 Evaluator Results

4.1 Results from the *Baseline* treatment of the *Evaluator Study*

Table 2 presents our main results on evaluators' beliefs, taken from the *Baseline* treatment of the *Evaluator Study*. Panel A presents the evaluators' prior beliefs in Column 1, overconfidence beliefs in Column 2, underconfidence beliefs in Column 3, implied Bayesian posterior beliefs in Column 4, and posterior beliefs in Column 5. Panel B presents evaluators' beliefs demeaned by the "truth." When considering evaluators' prior, implied Bayesian posterior, and posterior beliefs, we define the truth as the actual likelihood of a randomly-selected male/female worker having a poor performance. When considering evaluators' overconfidence and underconfidence beliefs, we define the truth as the actual likelihood of a randomly-selected male/female worker being overconfident conditional on having a poor performance and being underconfident conditional on having a good performance, respectively.[16] As a reference, we present the true values for male/female workers in the bottom rows of Table 2.

Column 1 ("Prior") of Table 2 shows the evaluators' prior beliefs—before they learn any information on workers' self-evaluations—about the likelihood that workers have poor performances. According to their priors, evaluators believe that female workers have a 42.95% chance of poor performances while male workers have a 39.08% chance of poor performances. That is, evaluators believe that female workers are 3.89 percentage points more likely to perform poorly. While this expected performance gap is statistically significant (Panel A), the expected performance gap is ultimately small and statistically indistinguishable from the true performance gap of 1.74 percentage points (Panel B). We summarize these findings in Result 1.

Result 1 (Prior Beliefs)**.** *According to evaluators' priors, the expected performance gap is small and statistically indistinguishable from the true performance gap.*

Column 2 ("Overconfidence") of Table 2 shows evaluators' beliefs about the likelihood that workers are overconfident. Evaluators believe men are much more likely to be overconfident: men are expected to be 8.25 percentage points significantly more likely than women to

[16] To calculate the chance of an evaluator being overconfident given a poor performance or underconfidence given a good performance, see Equations 4 and 5, respectively, in Appendix E.3.

Table 2: Evaluators' Beliefs in the *Baseline* treatment of the *Evaluator Study*

	Prior	Over-confidence	Under-confidence	Implied Bayesian Posterior	Posterior
	(1)	(2)	(3)	(4)	(5)
Panel A: Evaluators' Beliefs					
B(F)	42.97	39.86	55.68	43.83	61.85
B(M)	39.08	48.11	45.61	40.07	51.36
Δ	3.89**	-8.25***	10.07***	3.77**	10.49***
	(1.87)	(2.27)	(2.06)	(1.87)	(1.78)
Panel B: Evaluators' Beliefs - Truth					
B(F) - Truth(F)	-6.564	24.51	-19.12	-5.697	12.32
B(M) - Truth(M)	-8.710	9.051	-6.527	-7.725	3.572
Δ - Truth(Δ)	2.15	15.46***	-12.59***	2.03	8.75***
	(1.87)	(2.27)	(2.06)	(1.87)	(1.78)
N	402	402	402	402	402
Truth(F)	49.53	15.35	74.80	49.53	49.53
Truth(M)	47.79	39.06	52.14	47.79	47.79
Truth(Δ)	1.74	-23.70	22.65	1.74	1.74

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are robust and shown in parentheses. See Appendix Table A.2 for definitions of evaluators' beliefs. For the evaluator belief noted in the column, Panel A presents the average belief about female workers (see $B(F)$), the average belief about male workers (see $B(M)$), and the difference in these averages (see Δ). For the evaluator belief noted in the column, Panel B presents the average belief about female workers demeaned by the true value for female workers (see $B(F) - Truth(F)$), the average belief about male workers demeaned by the true value for male workers (see $B(M) - Truth(M)$), and the difference in these demeaned averages (see Δ - Truth(Δ)). At the bottom of the table, we provide corresponding true values for what evaluators' beliefs in Panel A should be if evaluators are fully accurate when they are asked to provide beliefs about female workers (see Truth(F)) or male workers (see Truth(M)) as well as the difference in these values (see Truth(Δ)). Data are from the 402 participants in the *Baseline* treatment of *Evaluator Study*.

believe that they have a good performance when considering workers who actually have a poor performance (Panel A). Nonetheless, this expected gender gap in overconfidence is significantly underestimated by 15.46 percentage points (Panel B).

Column 3 ("Underconfidence") of Table 2 shows the evaluators' beliefs about the likelihood that workers are underconfident. Evaluators believe women are much more likely to be underconfident: women are expected to be 10.07 percentage points significantly more likely than men to believe they have a poor performance when considering workers who actually have a good performance (Panel A). Nonetheless, this expected gender gap in underconfidence is significantly underestimated by 12.59 percentage points (Panel B).

We summarize these beliefs about overconfidence and underconfidence in Result 2, and

we provide additional evidence of the confidence gap being expected in Section 5.1.

Result 2 (Overconfidence and Underconfidence Beliefs)**.** *According to evaluators' overconfidence and underconfidence beliefs, the expected gender gaps in confidence are substantial and statistically significant but are smaller than the true confidence gaps.*

As explained in Section 2.3 (and detailed in Appendix E), we can use the three evaluator beliefs discussed so far to calculate evaluators' implied Bayesian posterior beliefs, i.e., what Bayesian evaluators would believe is the likelihood that a male/female worker has a poor performance after they are informed that 56%/80% of male/female workers believed that they had a poor performance. Column 4 ("Implied Bayesian Posteriors") of Table 2 presents these estimates. According to Bayesian updating, evaluators should expect that female workers are 3.77 percentage points more likely to have a poor performance than male workers—an expected performance gap that is statistically significant (Panel A) but small and statistically indistinguishable from the true performance gap of 1.74 percentage points (Panel B). That is, evaluators' implied Bayesian posterior beliefs indicate that the confidence gap should *not* be contagious: being provided with information on the workers' self-evaluations should *not* cause them to form overly pessimistic beliefs about women. As detailed in Appendix Section E.4, this results from the fact that, in our data, evaluators believe that workers are sufficiently miscalibrated in their self-evaluations such that a Bayesian evaluator would update very little from this information. We summarize these findings in Result 3.

Result 3 (Implied Bayesian Posterior Beliefs)**.** *According to evaluators' implied Bayesian posterior beliefs, the confidence gap should not be contagious. Thus, if evaluators are Bayesian, the expected performance gap—after being provided with information on workers' self-evaluations—should be small and statistically indistinguishable from the true performance gap.*

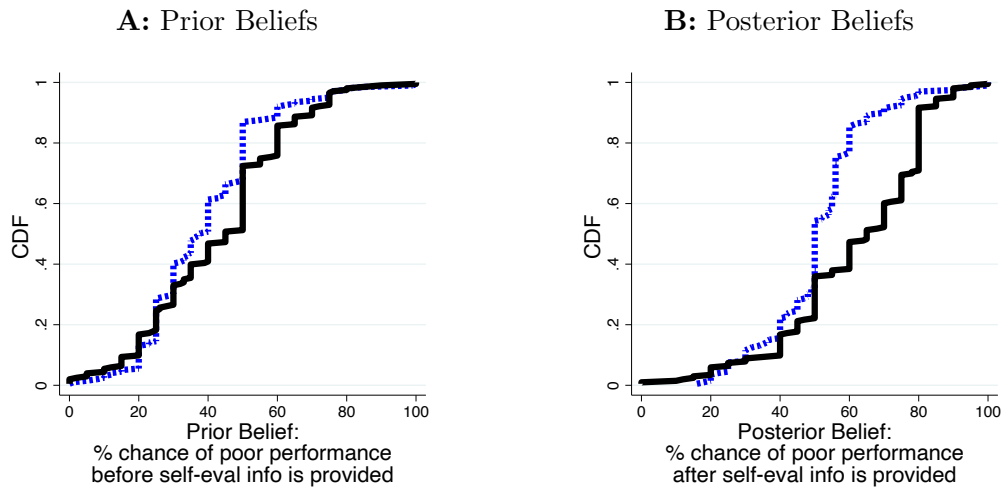
This proves not to be the case. The confidence gap proves to be contagious. Unlike their prior beliefs and unlike their implied Bayesian posterior beliefs, evaluators' posterior beliefs do not reflect a small to nonexistent expected performance gap. Rather, after learning about more optimistic self-evaluations from male workers or more pessimistic self-evaluations from female workers, evaluators expect a substantial and statistically significant performance gap. Specifically, Column 5 ("Posteriors") of Table 2 presents evaluators' posterior beliefs. Evaluators believe that female workers have a 61.85% chance of poor performance while male workers have a 51.63% chance of poor performance. That is, evaluators believe that female workers are 10.49 percentage points more likely to perform poorly. This expected performance gap is both statistically significant (Panel A) and substantially larger than the true performance gap of 1.74 percentage points (Panel B). Indeed, this expected performance

gap is more than 8.75 percentage points significantly larger than—or more than 6 times larger than—the true performance gap. In addition, when comparing priors to posteriors, the expected performance gap significantly increases by 6.61 percentage points.[17] We summarize these findings in Result 4.

Result 4 (Posterior Beliefs)**.** *According to evaluators' posterior beliefs, the confidence gap is contagious. The expected performance gap—after being provided with information on workers' self-evaluations—is substantial and significantly larger from the true performance gap.*

In summary, the confidence gap—conveyed via the gender gap in self-evaluations—exacerbates the expected performance gap, even though it should not if evaluators were Bayesians and even though evaluators expect a confidence gap (more on this in Section 5.1). This contagious confidence gap results in overly pessimistic beliefs about women relative to men, as also evident by the distributions of prior beliefs and posterior beliefs shown in Figure 1 (see also Appendix Figure B.1 for histograms).

Figure 1: Evaluators' Beliefs in *Baseline* treatment



Graphs show CDFs of the noted evaluators' beliefs from the *Baseline* treatment of the *Evaluator Study*.

4.2 Results from *Attention* and *Calculation* treatments: Testing interventions to help evaluators account for the confidence gap

One hypothesis as to why evaluators fail to accurately account for the confidence gap relates to attention. For instance, since evaluators' overconfidence and underconfidence beliefs do

[17] This 6.61 percentage point increase is statistically significant ($p < 0.01$) when regressing *prior-posterior* on an indicator for beliefs about female workers, with robust SEs.

reveal an expected gender difference in confidence, it could be that evaluators are simply *inattentive* to—but not unaware of—the influence of gender in self-evaluations when providing their posterior beliefs. This hypothesis could be enabled by the fact that we elicit evaluators' overconfidence and underconfidence beliefs only *after* they provide their posterior beliefs in the *Baseline* treatment. Thus, to investigate the role of attention, we turn to the *Attention* treatment. Unlike the *Baseline* treatment, the *Attention* treatment elicits evaluators' overconfidence and underconfidence beliefs *before* they provide their posterior beliefs, which may increase the evaluators' attention to a worker's confidence given their gender.

Table 3 directly compares the *Baseline* treatment to the *Attention* treatment and shows that—for all evaluator beliefs—the expected performance gap is not significantly different between the *Baseline* treatment and *Attention* treatment.[18] The coefficient estimates on Δ reproduce the expected performance gap in the *Baseline* treatment, while the coefficient estimates on Δ***Attention* show how the expected performance gap changes in the *Attention* treatment relative to the *Baseline* treatment. The coefficient estimates on Δ***Attention* are small and never statistically significant. Thus, our treatment manipulation designed to encourage evaluators to think about a worker's confidence given the worker's gender before reporting their posterior beliefs does not significantly reduce the extent to which evaluators' posterior beliefs indicate an expected performance gap. We summarize these findings in Result 5.

Result 5 (Impact of Attention Treatment)**.** *The contagious confidence gap is not mitigated by eliciting evaluators' overconfidence and underconfidence beliefs before they provide their posterior beliefs. According to evaluators' posteriors in the Attention treatment, the expected performance gap remains substantial and significantly larger than the true performance gap.*

Another hypothesis for evaluators' failure to accurately account for the confidence gap is that—perhaps due to the cognitive difficulty of Bayesian updating—they may struggle to accurately update their priors in response to the signal about worker performance that is conveyed via their self-evaluations. Thus, to investigate the effectiveness of a more extreme intervention that may help evaluators overcome any difficulty with Bayesian updating, we turn to the *Calculation* treatment. Like the *Attention* treatment, the *Calculation* treatment elicits evaluators' overconfidence and underconfidence beliefs *before* eliciting their posterior beliefs. In addition, the *Calculation* treatment uses evaluators' overconfidence and underconfidence beliefs—along with their prior beliefs—to inform evaluators of their implied Bayesian posteriors before they provide their posterior beliefs.

[18] Following the structure of Table 2, Appendix Table B.2 presents the results for the *Attention* treatment. Evaluators' beliefs in the *Attention* treatment are very similar to those in the *Baseline* treatment.

Table 3 directly compares the *Baseline* treatment to the *Calculation* treatment and reveals one set of significant differences.[19] According to evaluators' posteriors, the expected performance gap is significantly smaller in the *Calculation* treatment than in the *Baseline* treatment (Column 5 of Panel A) and significantly more accurate in the *Calculation* treatment than in the *Baseline* treatment (Column 5 of Panel B), while there are no significant differences in other beliefs. Thus, helping evaluators to update in a Bayesian manner in response to information on workers' self-evaluations significantly reduces the extent to which evaluators expect a performance gap. We summarize these findings in Result 6.

Result 6 (Impact of Calculation Condition)**.** *The contagious confidence gap is somewhat mitigated by informing evaluators of their implied Bayesian posterior beliefs before they provide their posterior beliefs. According to evaluators' posteriors in the Calculation treatment, the expected performance gap becomes substantially smaller and more accurate in the Calculation treatment when compared to the expected performance gap in the Baseline treatment.*

Given the effectiveness of the *Calculation* treatment, a natural question relates to the extent to which the *Calculation* treatment induces a sort of experimenter demand effect or social desirability bias. Indeed, it may be the case that social pressure—whether from the experimenter, colleagues, or others—is a crucial component in encouraging individuals to accurately account for gender differences in confidence. It may also be the case that teaching individuals about Bayesian updating is somewhat inseparable from conveying to individuals how they *should* form their beliefs. Thus, while this type of experimenter demand effect or "teaching" could contribute to the results in the *Calculation* treatment, we leave open the possibility that this is a feature, not a bug.

Regardless, we provide three pieces of evidence that point against the relevance of experimenter demand effects or social desirability bias in our *Calculation* treatment. First, the majority of participants (61%) in the *Calculation* treatment report a posterior belief that differs from their implied Bayesian posterior belief, which shows that most participants are not simply reporting back the number that is suggested to them. Second, our results hold when only considering this 61% of participants.[20] Third, as will become evident in Section 4.3, we will be able to show that—to the extent experimenter demand effects or social desirability

[19]Following the structure of Table 2, Appendix Table B.3 presents the results for the *Calculation* treatment. Evaluators' beliefs in the *Calculation* treatment are similar to those in the *Baseline* and *Attention* treatment with one notable exception. While evaluators' posterior beliefs in the *Calculation* treatment indicate that they expect a performance gap, this expected performance gap is only *marginally* significantly different than the true gap and is noticeably smaller than what was observed in the other two treatments.

[20]For the 61% of evaluators with differing posterior and implied Bayesian posterior beliefs in the *Calculation* treatment, evaluators' posterior beliefs indicate an expected performance gap of 6.37 percentage points, which remains smaller than the expected gap of 10.49 percentage points in the *Baseline* treatment.

Table 3: Evaluators' Beliefs in the *Baseline, Attention, and Calculation* treatment of the *Evaluator Study*

	Prior	Over-confidence	Under-confidence	Implied Bayesian Posterior	Posterior
	(1)	(2)	(3)	(4)	(5)
Panel A: Evaluators' Beliefs					
Δ	3.89**	-8.25***	10.07***	3.77**	10.49***
	(1.87)	(2.27)	(2.06)	(1.87)	(1.78)
Δ*Attention	-0.47	3.65	-0.23	-0.23	0.36
	(2.62)	(3.16)	(2.93)	(2.60)	(2.48)
Δ*Calculation	-0.81	-1.17	1.66	-0.66	-5.57**
	(2.61)	(3.21)	(2.86)	(2.56)	(2.54)
Panel B: Evaluators' Beliefs - Truth					
Δ	2.15	15.46***	-12.59***	2.03	8.75***
	(1.87)	(2.27)	(2.06)	(1.87)	(1.78)
Δ*Attention	-0.47	3.65	-0.23	-0.23	0.36
	(2.62)	(3.16)	(2.93)	(2.60)	(2.48)
Δ*Calculation	-0.81	-1.17	1.66	-0.66	-5.57**
	(2.61)	(3.21)	(2.86)	(2.56)	(2.54)
	(1.83)	(2.20)	(2.08)	(1.80)	(1.73)
N	1210	1210	1210	1209	1210
Condition FE	yes	yes	yes	yes	yes
Truth(Δ)	1.74	-23.70	22.65	1.74	1.74

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are robust and shown in parentheses. See Appendix Table A.2 for definitions of evaluators' beliefs. For the type of evaluator belief noted in the column, Panel A presents an OLS of evaluators' belief on (i) suppressed indicators (i.e., Condition FEs) for the *Baseline* treatment, the *Attention* treatment, and the *Calculation* treatment, as well as (ii) an indicator for being asked about female workers (Δ), an indicator for being asked about female workers interacted with the indicator for the *Attention* treatment (Δ*Attention), and an indicator for being asked about female workers interacted with the indicator for the *Calculation* treatment (Δ*Calculation). For the type of evaluator belief noted in the column, Panel B presents an OLS of evaluators' beliefs demeaned by the true values on the same set of indicators as in Panel A. At the bottom of the table, we provide corresponding true values for the difference in evaluators' beliefs about female and male workers if evaluators are fully accurate when they are asked about female and male workers (see the estimates Truth(Δ)). Data are from the 1210 participants in the *Baseline, Attention,* or *Calculation* treatment of *Evaluator Study*. Sample size differs slightly in column (4) as some evaluators' beliefs imply a Bayesian posterior that is undefined.

bias drive the effectiveness of the *Calculation* treatment—this is not specific to gender (i.e., it is not specific to a potentially sensitive topic). Even in the *Unknown Gender* treatments, the *Calculation* treatment proves effective.

In more broadly considering our results, there are three other reasons as to why we are

even less worried about experimenter demand effects or social desirability bias in our other results, including our main results from the *Baseline* treatment. First, across our treatments, the type of experimenter demand effects or social desirability bias that one may naturally be most concerned about is that people do not want to appear biased against women. To the extent this is the case, we note that it would work opposite our results—i.e., it would make it more difficult for us to find posterior beliefs that are biased against women. Second, the desire to avoid experimenter demand effects or social desirability bias motivated why we only ever ask evaluators about male workers or female workers (although, as discussed in Section 6.5, we later investigate the impact of simultaneously asking about male and female workers). Third, while the *Calculation* treatment involves an extensive intervention, the design of the *Attention* treatment seeks to subtly increase evaluators' attentiveness to gender differences in confidence simply by changing the order in which beliefs are elicited.

4.3 Results from *Unknown Gender* treatments: Testing for underlying problems in updating that are not specific to gender

To narrow in on how much beliefs about gender drive our results, we now turn to the *Unknown Gender* treatments of the *Evaluator* Study. In these treatments, evaluators are either asked about group-1 (male) workers or group-2 (female) workers. While evaluators know that these groups reflect how workers answered a follow-up question, evaluators are never provided with any demographic information (i.e., gender) on these groups.

Following the structure of Table 2, Appendix Tables B.4–B.6 separately present the results from each of the three *Unknown Gender* treatments. There are three main takeaways. First, according to their prior beliefs and as one would expect given the lack of information provided about group-1 and group-2 workers, evaluators in each treatment do not expect a performance gap. Second, evaluators in each treatment directionally, and sometimes to a statistically significant degree, expect that group-1 (male) workers are more likely to be overconfident conditional on a poor performance and that group-2 (female) workers are more likely to be underconfident conditional on a good performance. This demonstrates that—even without information on gender—evaluators quite reasonably believe a group of workers is relatively more underconfident and relatively less overconfident when they learn that 80% of workers in that group expect a poor performance compared to when they learn that 56% of workers in that group expect a poor performance. Third, the confidence gap again results in overly pessimistic beliefs about women relative to men: according to their posterior beliefs, evaluators in each treatment expect that group-2 (female) workers are significantly more

likely to have a poor performance.[21]

Indeed, the extent to which the confidence gap results in overly pessimistic beliefs about women is similar when the worker's gender is and is not known. To see this, each column in Table 4 presents posterior beliefs from a pair of treatments that compares the X treatment and the *X, Unknown Gender* treatment for $X \in$ {Baseline, Attention, Calculation}.

Table 4: Evaluators' Posterior Beliefs about Workers according to whether or not they are in a *Unknown Gender* treatment of the *Evaluator Study*

	X and *X, Unknown Gender* Condition Given $X =$		
	Baseline	*Attention*	*Calculation*
	(1)	(2)	(3)
Panel A: Evaluators' Beliefs			
Δ	10.49***	10.85***	4.92***
	(1.78)	(1.73)	(1.81)
Δ*Unknown Gender	0.57	-0.29	-0.05
	(2.40)	(2.45)	(2.53)
Panel B: Evaluators' Beliefs - Truth			
Δ	8.75***	9.11***	3.18*
	(1.78)	(1.73)	(1.81)
Δ*Unknown Gender	0.57	-0.29	-0.05
	(2.40)	(2.45)	(2.53)
N	807	795	798
Condition FE	yes	yes	yes
Truth(Δ)	1.74	1.74	1.74

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are robust and shown in parentheses. The data are from the X and *X, Unknown Gender* treatments noted in the columns. Panel A presents an OLS of evaluators' posterior beliefs on (i) suppressed indicators (i.e., Condition FEs) for the X treatment and the corresponding *X, Unknown Gender* treatment as well as (ii) an indicator for being asked about female workers (Δ) and an indicator for being asked about female workers interacted with the indicator for the *X, Unknown Gender* treatment (Δ*Unknown Gender). Panel B presents an OLS of evaluators' posterior beliefs demeaned by the true values on the same set of indicators as in Panel A. At the bottom of the table, we provide corresponding true values for the difference in evaluators' beliefs about female and male workers if evaluators are fully accurate when they are asked about female and male workers (see the estimates Truth(Δ)). Data are from the 2400 participants in *Evaluator Study*, split across the three columns according to the relevant treatments.

Across all three pairs of treatments, Table 4 reeals no significant differences in posteriors in the *Unknown Gender* treatments compared to those where gender is known (see the

[21] That said, akin to the findings when gender is known, evaluators' posteriors indicate an expected performance gap between these two groups that is smaller and more accurate in the *Calculation, Unknown Gender*. Also, see Appendix Table B.7 to compare the beliefs across the three *Unknown Gender* treatments.

coefficients on Δ*Unknown Gender).[22] Thus, we do not find any evidence for evaluators accounting, to even a small degree, for the *gender* gap in self-evaluations. Rather, the confidence gap in self-evaluations—regardless as to whether the gender of those providing the self-evaluations is known—results in a substantial expected performance gap between men and women. Interestingly, since these results suggest that knowing gender does not help evaluators to account for the confidence gap, these results could contribute to the potentially limited efficacy of certain policy interventions, such as gender-blind reviews designed to reduce gender discrimination (Kolev et al., 2019). We summarize these findings in Result 7.

Result 7 (Impact of Known Gender Being Known)**.** *According to evaluators' posteriors, consistent with them not accounting for the role of gender, the expected performance gap remains just as large when the gender of workers is unknown rather than known.*

5 Heterogeneity

Evaluators expect a significant confidence gap, and their implied Bayesian posterior beliefs suggest that they should be able to (almost entirely) account for this. Nonetheless, evaluators do not account for the confidence gap as conveyed via workers' self-evaluations, and instead it leads to a large expected performance gap. Put differently, this contagious confidence gap results in substantial gender disparities. It is therefore important to understand whether certain types of evaluators drive this bias (our focus below) and to assess the robustness of our results (our focus in Section 6). Below, to investigate whether certain types of evaluators drive our results, we present a series of heterogeneity analyses. For conciseness, we will focus on evaluators' posterior beliefs from the *Baseline* treatment—while also showing how posterior beliefs are similar in the *Attention* treatment and indicate less of an expected difference between men and women in the *Calculation* treatment.

5.1 Do our results persist for evaluators who expect gender differences in confidence?

One could worry that our confidence elicitation is complicated or noisy or otherwise does not capture evaluators' true expectations about gender differences in confidence.[23] To provide

[22] If we consider evaluators' other beliefs, only two small differences arise. First, while evaluators' priors (and sometimes their posteriors) indicate that they expect a small performance gap when the worker gender is known, this is no longer the case when worker gender is unknown. Second, while evaluators' confidence beliefs indicate that they expect men to be significantly more overconfident and women to be more significantly more underconfident when worker gender is known, this is less true when worker gender is unknown.

[23] While we did not directly elicit confidence in one's beliefs, we find that over/underconfidence beliefs typically do not fall at 50% (see the histograms in Appendix Figure B.2), which might have been an indicator

additional evidence of the confidence gap being expected—and our results persisting among evaluators who expect the confidence gap—we can turn to data from two follow-up survey questions and to data from one of our additional study versions.

The two follow-up survey questions directly ask evaluators to categorize the relative confidence of men versus women. The first question asks evaluators to categorize the relative confidence of men versus women in general. While 46% of evaluators expect no gender difference in confidence, nearly all of remaining evaluators expect the confidence gap: 51% believe that women are less confident but only 3% believe that men are less confident. The second question asks specifically about confidence in math and science tasks, and similar results follow: while 42% of evaluators expect no gender difference in confidence, 51% believe that women are less confident while only 7% believe that men are less confident.

Appendix Table C.1 reproduces Column 5 of Table 3 for each of these groups of evaluators. These results reveal that even evaluators who think women are less confident than men (Columns 1 and 4) fail to account for the confidence gap: their posterior beliefs reveal a substantial and statistically significant expected performance gap. Similar results hold among evaluators who think there is no gender difference in confidence (Columns 2 and 5). The results are noisier when restricting to the group of evaluators who think women are more confident than men (Columns 3 and 6), likely due to the small sample size of this group.

In summary, most evaluators think that women are less confident than men, and almost no evaluators think the reverse is true. Related, our results persist even when we only consider evaluators who directly say that there is a confidence gap. In addition, as shown in Appendix Table D.9 and as discussed in Appendix D.4, we can show—in a different study version in which we elicit evaluators' confidence beliefs about both men and women—that our results persist among evaluators with incentivized overconfidence beliefs that directly indicate that they believe men are more overconfident than women, and among evaluators with incentivized underconfidence beliefs that directly indicate that they believe women are more underconfident than men.

5.2 Do our results persist for evaluators who think they accurately accounted for the gender gap in self-evaluations?

One might suspect that evaluators—if prompted to reflect on it—are aware that they did or did not accurately account for the gender gap in self-evaluations in our study. To investigate this, we can turn to data from the following question that we ask in in the follow-up survey of the known gender treatments: "When providing your predictions in this study, to what

of evaluators being entirely unsure about the confidence of men and women.

extent were you accounting for any gender differences in confidence?" 63% of evaluators answer "neither too little nor too much," 14% of evaluators answer "slightly or far too much," and 23% of evaluators answer "far or slightly too little."

Appendix Table C.2 reproduces Column 5 of Table 3 for each group of evaluators. Each group of evaluators expects a performance gap, according to their posterior beliefs. In addition, the expected performance gap is the *smallest* among evaluators who believe that they adjusted *too little* for gender differences in confidence. Finally, when we instead ask evaluators whether they think employers—rather than themselves—accurately account for the confidence gap, similar results follow (see Appendix Table C.3).

5.3 Are our results driven by evaluators who exhibit other cognitive biases?

It could be the case that our results are driven, in part, by cognitive biases or errors. To investigate this, we incentivize evaluators to correctly answer five additional questions at the end of the study: a standard Bayesian updating question, a question designed to detect base rate neglect (Kahneman and Tversky, 1972), and the three-question cognitive reflection test (CRT) (Frederick, 2005), all presented in random order.[24]

Appendix Table C.4 presents results on how these measures correlate with the extent to which evaluators expect a performance gap, according to their posterior beliefs. Counter to cognitive errors or updating failures explaining our results, the expected performance gap is directionally larger for evaluators with higher cognitive ability scores (Column 1) and is directionally smaller for evaluators who give a response farther from the Bayesian posterior in the Bayesian updating question (Column 4). But, consistent with base rate neglect contributing to our results, the expected performance gap is directionally larger—sometimes significantly so—for evaluators who exhibit pure base rate neglect (Column 2) or who give a response farther from the Bayesian posterior in the base rate neglect question (Column 3).[25] This, together with the previously-discussed *Calculation* treatment results, suggests that there may be large gains in equality by helping to alleviate some cognitive biases, as these biases are correlated with the failure to account for the confidence gap.

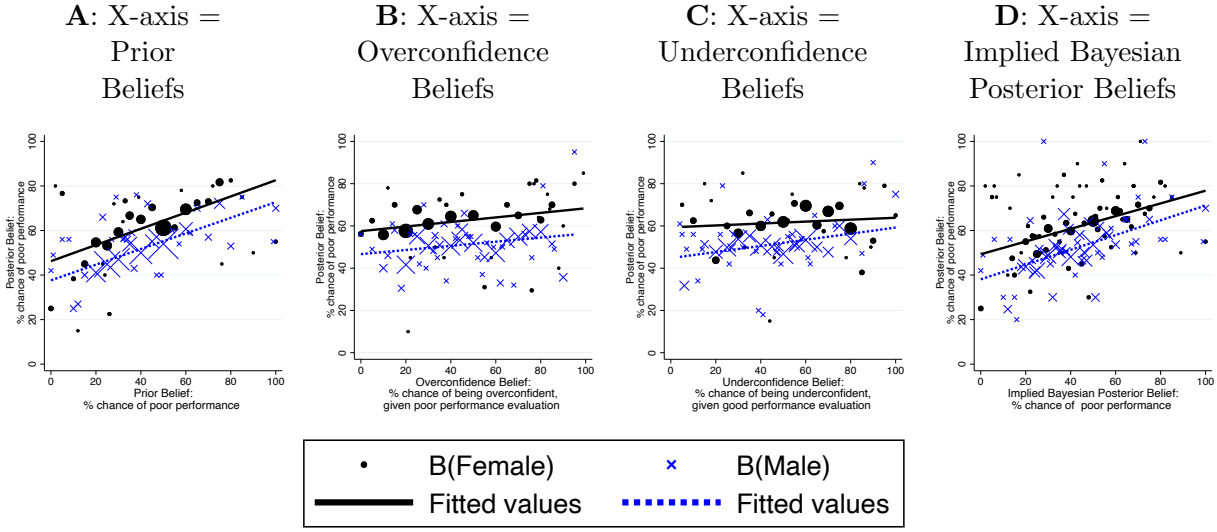
[24]For full question text, see supplemental Online Appendix Figures F.2.12–F.2.16.

[25]Consistent with the *Calculation* treatment helping to eliminate the role of cognitive biases, we find that these relationships weaken in the *Calculation* treatment.

5.4 Are our results driven by evaluators with certain other beliefs?

One may wonder whether our results are driven by evaluators who hold a particular set of initial beliefs. For instance, perhaps evaluators who seem most unsure about the chance that a male or female worker has a poor performance—and hence report a prior belief of 50%—are more susceptible to being influenced by information on workers' self-evaluations. This proves not to be the case (and we further note that only around 20% of evaluators have prior beliefs that fall right at 50%, as shown in Appendix Figure B.1). For evaluators in the *Baseline* treatment, Figure 2 plots posterior beliefs as a function of evaluators' prior beliefs (Panel A), overconfidence beliefs (Panel B), underconfidence beliefs (Panel C), and implied Bayesian posterior beliefs (Panel D). These results make clear that evaluators' posterior beliefs disfavor women relative to men across the entire range of evaluators' other beliefs.[26]

Figure 2: *Baseline Treatment*: Posterior Beliefs as a Function of Their Other Beliefs



Graphs show a scatter plot (dots weighted by sample size) of evaluators' posterior beliefs as a function of their beliefs noted on the horizontal axis. Data are from the *Baseline* treatment of the *Evaluator Study*.

5.5 Are our results driven by evaluators with certain demographic characteristics?

To investigate if evaluators' posterior beliefs vary by demographics, we reproduce Column 5 of Table 3 for various demographic groups. Specifically, Appendix Table C.5 splits evaluators

[26] Appendix Figure C.1 shows that similar results follow in the *Attention* treatment. Appendix Figure C.2 shows that evaluators' prior beliefs and implied Bayesian posterior beliefs are more predictive in the *Calculation* treatment, which is perhaps related to the smaller expected performance gap in that treatment.

according to their gender, educational attainment, and income, while Appendix Table C.6 splits evaluators according to their age and political affiliation.

As evidence against the hypothesis that some demographic groups (e.g., women, or more educated participants) are better-able to account for the confidence gap, all subgroups of evaluators have posterior beliefs that significantly disfavor women.

6 Robustness

In Section 6, to investigate the robustness of our results, we turn to additional study versions that involve various design changes to the *Baseline* treatment of the *Evaluator Study*.

6.1 Are our results robust to evaluators with hiring and managerial experience providing beliefs about typical job candidates?

One may wonder whether evaluators could do a better job of accounting for the confidence gap if they had more hiring and managerial experience and if they were asked about men and women who may be more "typical" of likely job candidates. To investigate this, we ran the *Baseline* treatment and *Baseline, Unknown Gender* treatment of the *Evaluator (Professional Evaluators) Study* (see Appendix Section D.1 for an overview and the supplemental Online Appendix F.6 for full experimental instructions).

Specifically, we recruited 800 professional participants who—according to self-reported data—met the following two criteria: (1) they have experience in making hiring decisions (i.e. have been responsible for hiring job candidates) and (2) they have experience in a management position.[27] In addition, rather than asking them about male and female workers recruited from Prolific, we asked them about people who are likely to be applying for jobs in the near future: male and female workers who are undergraduate students at a large Midwestern university and expect to graduate in 2023.[28]

Following Table 1, Appendix Table D.1 presents the results for these undergraduate students and confirms that the confidence gap persists for them. Despite an insignificant performance gap of 1.91 percentage points, there is a substantial and statistically significant

[27]Specifically, we use the internal screening questions on Prolific to recruit this sample. Participants answers to these questions are self-reported, and we cannot verify their work experience. That said, we note that the vast majority of Prolific participants do not meet these screening restrictions and that recent other papers who have used similar approaches include Huber and Huber (2020) and Saccardo and Serra-Garcia (2022). In our own follow-up survey, we can also confirm that 81% of these participants responded "yes" when asked a different but similar question to Prolific screeners – i.e., when asked "Do you have any experience with decisions that relate to the hiring, pay, or promotion of employees or fellow colleagues?"

[28]Specifically, we recruited these participants through Ohio State University in March/April of 2022. See Appendix Table A.6 for more details, and Online Appendix F.1 for full experimental instructions.

confidence gap of 26.3 percentage points: 58.6% of female workers believe they have a poor performance while only 32.3% of male workers believe they have a poor performance.

Following Table 2, Appendix Table D.2 presents the results for the professional evaluators. According to their priors, professional evaluators appear slightly more accurate than our main evaluators in terms of their expected performance gap. As shown in Column 1, professional evaluators expect an insignificant performance gap of 1.86 percentage points (Panel A), which is nearly identical to the true performance gap of 1.91 percentage points (Panel B). In addition, professional evaluators expect a confidence gap (Columns 2 and 3) and—if they are Bayesian—should not be influenced by the confidence gap (Column 4). Nonetheless, just as with our main participants, the confidence gap proves contagious and causes professional evaluators to form overly pessimistic beliefs about women relative to men. According to their posteriors, professional evaluators inaccurately expect a substantial and statistically significant performance gap of 14.65 percentage points (Column 5).

In addition, comparing professional evaluators' beliefs in the *Baseline* treatment to professional evaluators' beliefs in the *Baseline, Unknown Gender* treatment further confirms that we observe no evidence for professional evaluators accounting for *gender* differences when forming their posterior beliefs.[29]

Taken together, despite some evidence for the expected performance gap being more accurate according to the professional evaluators' priors beliefs, professional evaluators appear just as poor at accounting for the confidence gap when forming posterior beliefs in response to information on workers' self-evaluations.

6.2 Are our results robust to evaluators gaining more experience with worker self-evaluations?

One may wonder whether evaluators could do a better job of accounting for confidence if they had more experience with the exact type of self-evaluations in our study. To investigate this, we ran the *Baseline* treatment of the *Evaluator (Extended) Study* (see Appendix Section D.2 for an overview and the supplemental Online Appendix F.3 for full experimental instructions).

In this study, prior to providing their posterior beliefs about a male or female worker who is randomly selected from the available pool of workers, evaluators are asked to make 20 predictions about specific workers after learning each worker's self-evaluation. As shown in Appendix Table D.4, gaining experience with self-evaluations does not help evaluators

[29]The expected gender gap in performance according to evaluators' posterior beliefs—i.e., the estimates on Δ in Columns 5 of Appendix Tables D.2 and D.3—are statistically indistinguishable ($p > 0.1$).

to better account for the gender gap in confidence. According to their posteriors, even experienced evaluators expect a large and statistically significant performance gap (∼15 percentage points).

6.3 Are our results robust to beliefs about specific workers?

To investigate if our results are robust to evaluator beliefs that pertain to a specific worker–after learning only that worker's self-evaluation—we turn to the worker-specific beliefs that evaluators provide in the *Baseline* treatment of the *Evaluator (Extended) Study*, described above in Section 6.2 (again, see Appendix Section D.2 for an overview and the supplemental Online Appendix F.3 for full experimental instructions).

As shown in the northeastern region of Appendix Figure D.2, there is some evidence that evaluators account for the confidence gap among the most pessimistic self-evaluations. For instance, when a worker reports an 80% chance of having a poor performance in their self-evaluation, the average evaluator believes there is a 74% chance of that worker having a poor evaluation if the worker is a man but only a 70% of that worker having a poor evaluation if that worker is a woman. Nonetheless, Appendix Table D.5 shows that—even when asked about specific workers—evaluators expect a statistically significant performance gap (∼4.65 percentage points), according to their posterior beliefs.

6.4 Do our results persist when workers face strategic incentives?

In the main *Worker Study*, workers are incentivized to report accurate beliefs. However, in many settings outside the lab, individuals might have incentives to strategically inflate their self-reported beliefs in order to promote themselves to potential evaluators (e.g., teachers, colleagues, supervisors, employers, clients, etc.). We ran the *Strategic Incentives* treatment of the *Worker Study* and the *Strategic Incentives* treatment of the *Evaluator (Extended) Study* to investigate whether evaluators are more (or less) inclined to adjust for the confidence gap when they know that workers face strategic incentives (see Appendix Section D.3 for an overview and the supplemental Online Appendices F.1 and F.3 for full experimental instructions).

In the *Strategic Incentives* treatment of the *Worker Study*, workers face strategic incentives to inflate their performance since they earn more money if they hired are by an "employer" who learns their self-evaluation. In the *Strategic Incentives* treatment of the *Evaluator (Extended) Study*, evaluators are provided with the self-evaluations of these workers, and are informed of the workers' strategic incentives.

Following Table 1, Appendix Table D.6 presents the results for workers who face these

strategic incentives. Despite an insignificant performance gap of 1.44 percentage points among the available pool of workers, there is a substantial and statistically significant confidence gap of 17 percentage points: 74% of female workers believe they have a poor performance while only 57% of male workers believe they have a poor performance.

Following Table 2, Appendix Table D.7 presents the results for evaluators who are asked about these workers. Even these evaluators expect a large and statistically significant performance gap (∼9 percentage points), according to their posterior beliefs.

6.5 Are our results robust to being asked about both men and women?

In the evaluator results discussed so far, evaluators provide beliefs about only the group of male workers or only the group of female workers. Building off of prior work that suggests judgments are less reasoned when comparison information is lacking, Bohnet et al. (2016) find that evaluators are less likely to be influenced by group stereotypes when simultaneously reviewing two resumes (one for a man and one for a woman) than when reviewing resumes one at a time. Inspired by this finding, we ran the *Joint Evaluations* treatment of the *Evaluator (Extended) Study* (see Appendix Section D.4 for an overview and the supplemental Online Appendix F.3 for details). Specifically, in the *Joint Evaluations* treatment, we investigate whether evaluators are better able to account for the confidence gap when they are asked—on the same decision screen—to provide beliefs about a male worker and a female worker.

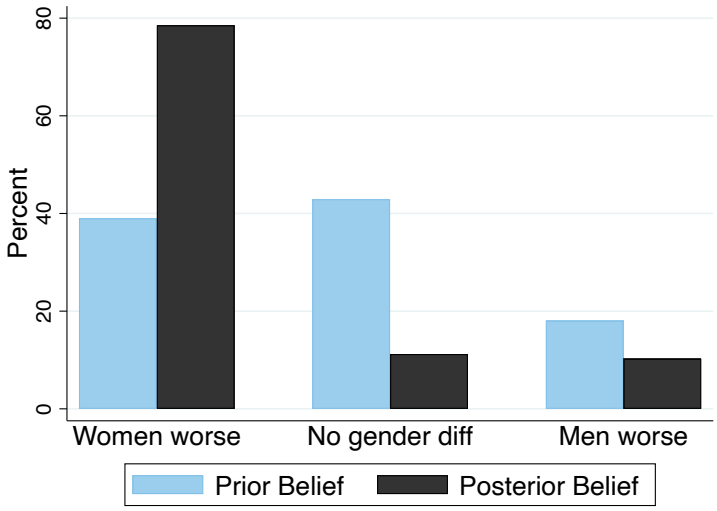
Following the same specifications as those in Table 2, Appendix Table D.8 presents results for the evaluators in the *Joint Evaluations* treatment. Joint evaluations do not eliminate the expected performance gap: even these evaluators expect a large and statistically significant performance gap (∼15 percentage points), according to their posteriors.

6.6 Are our results robust to considering evaluators' beliefs at the individual-level?

Since our main evaluators provide beliefs about only one group of workers, our main evaluator results do not allow us to classify evaluators—at the individual-level—according to whether they expect female workers to be more, equally, or less likely to have a poor performance than male workers. But, our *Joint Evaluations* treatment allows for such classifications, which are presented in Figure 3.

When evaluators are classified according to their prior beliefs, shown via the light blue bars, we find that the percent of evaluators who think female workers are more, equally, or less

Figure 3: *Joint Evaluations Treatment*: Classifying Evaluators According to Their Beliefs



This graph shows the percent of evaluators who, given their prior or posterior beliefs, believe that women–relative to men—are more, equally, or less likely to have a poor performance in the first two, middle two, and right two bars, respectively. Data are from the *Joint Evaluations* treatment of the *Evaluator Study*.

likely to have a poor performance is 39%, 43%, and 18% respectively. But, the confidence gap causes a substantial increase—indeed a doubling—in the percent of evaluators who believe that female workers are more likely to have a poor performance than male workers. When evaluators are classified according to their posterior beliefs, shown via the black bars, the percent of evaluators who think female workers are more, equally, or less likely to have a poor performance is 79%, 11%, and 10% respectively. Thus, even when considering the individual-level results, the confidence gap is contagious.

6.7 Are our results robust to evaluator beliefs when asked about other types of performance outcomes?

Our main results ask evaluators to provide beliefs about the likelihood that a worker has a poor performance. To assess the robustness of our evaluator results to beliefs about other types of performance outcomes, we ran an additional study called the *Evaluator (Alternative Questions) Study*. Compared to our main *Evaluator Study*, the main difference in this study is that, in addition to providing beliefs about the likelihood of a worker having poor performance in the manner defined in our main self-evaluation question (beliefs shown in Appendix Table A.2), evaluators are also asked to provide beliefs relating to five different performance outcomes—four of which involve objective outcomes (beliefs show in Appendix Table A.3; see Appendix Section D.5 for details).

Appendix Table D.11 presents the results for evaluator beliefs relating to these six performance outcomes. We confirm our main results directionally—and almost always to a statistically significant degree—across all six performance outcomes. To begin, for all six performance outcomes, evaluators' prior beliefs indicate little to no expected performance gap, evaluators' confidence beliefs indicate that they expect men to be more likely to be overconfident and women to be less likely to be underconfident, and evaluators' implied Bayesian beliefs indicate that learning about the workers' self-evaluations should not—if they are Bayesian—result in any expected gender difference in performance. Nonetheless, for five out of the six performance outcomes, evaluators expect a large and statistically significant performance gap (∼5–10 percentage points), according to their posteriors.[30]

6.8 Are our results robust to conveying gender more subtly?

In all of the results discussed so far, evaluators are asked to make predictions about workers who are only labeled according to their gender. Since this design feature likely draws attention to gender, one might wonder whether experimenter demand effects contribute to our results. We expect the opposite is the case, since both increased attention to gender and social desirability bias would seem to point towards evaluators being *more* likely to accurately adjust for the gender gap in self-evaluations. Nonetheless, since individuals in many situations form beliefs about others after learning more than just their gender, an interesting question is whether our results persist when evaluators are informed of a worker's gender along with several other demographic characteristics. This is what we do in the *Evaluator (Additional Demographics) Study* (see Appendix Section D.6 for an overview and the supplemental Online Appendix F.5 for full experimental instructions).

In the *Evaluator (Additional Demographics) Study*, evaluators are told that their worker will be randomly drawn from a group of workers who work full time, are between 26 and 40 years old, live in the Southern region of the United States, have completed at least some college education, and are (wo)men.

Following Table 2, Appendix Table D.12 presents results for the evaluators in the *Evaluator (Additional Demographics) Study*. When evaluators provide beliefs about workers for whom gender information is more subtly conveyed, the expected performance gap remains: evaluators expect a large and statistically significant performance gap (∼23 percentage point), according to their posteriors. Thus, in addition to the aforementioned reasons, we observe no empirical evidence for our results being influenced by an experimenter demand effect

[30] Only when asked about the percent chance of participants getting 7+ question right is a gender difference *not* expected in posterior beliefs—and this lack of a gender difference aligns with the one case (see Column 6 of Table 1) in which male workers do not provide more confident self-evaluations than female workers.

regarding the salience of gender.

6.9 Are our results robust to situations where more information is known about the quality of worker?

One may wonder whether our results are robust to situations where more information is known about the quality of workers. Since our main self-evaluation question involves a subjective measure of performance, we can inform evaluators of a worker's *objective* performance and then investigate how the evaluators update about this *subjective* measure. This is what we do in the *Evaluator (Known Performance) Study* (see Appendix Section D.7 for an overview and supplemental Online Appendix F.5 for details).

Specifically, in the *Evaluator (Known Performance) Study*, evaluators are told that their worker will be randomly drawn from the group of male or female workers who got 5 questions right on the math and science test—ensuring their worker's *absolute performance* is known with certainty. Then, as in our main *Evaluator Study*, we elicit prior, posterior, and confidence beliefs about whether their worker has a poor performance.

Following Table 2, Appendix Table D.13 presents the results from the *Evaluator (Known Performance) Study*. Even when evaluators are given precise information on a worker's quality, evaluators expect a large and statistically significant performance gap ($\sim$14 percentage points), according to their posterior beliefs. Thus, our main results are robust to precise information on a worker's quality being known. Furthermore, our main results are robust to situations where there is neither an actual nor believed gender difference in performance.

7 Discussion

Through a series of experiments, we document that evaluators *expect* a confidence gap, but they do not *account for* it. Specifically, we show that the confidence gap—conveyed via workers' self-evaluations—results in overly pessimistic beliefs about women relative to equally-performing men. This "contagious" confidence gap arises even though it should not have if evaluators were Bayesian, and even when evaluators expect the confidence gap. This contagious confidence gap is indeed pervasive across subsamples and is robust to many features of the environment. Only a targeted intervention that helps evaluators with Bayesian updating proves somewhat effective at eliminating the expected performance gap.

We see many important avenues for future work. First, our results suggest that—to eliminate confidence-driven gender gaps in hiring, promotion, and pay decisions—it need not be enough to increase the *awareness* of the confidence gap. Instead, people may need

help taking steps to account for it. Future work may investigate the effectiveness of such steps. For instance, motivated by the findings from the *Calculation* treatment, future work may investigate whether—rather than simply asking employers to provide their final beliefs about the quality of candidates—it is helpful to ask employers about various inputs into those final beliefs and then provide guidance on how to form final beliefs given those inputs.

Second, our results do not rule out the possibility that—in some environments—increasing the awareness of the confidence gap may prove to be a useful policy intervention. Thus, future work should investigate when and where increasing the awareness of the confidence gap proves beneficial, particularly in more complex environments in which attention to the confidence gap may be muted, such as those involving more free-form communication (Coffman et al., 2019b) or those that require updating from the lack of information (Enke, 2017; Charness et al., 2022).

Third, given the effectiveness of our *Calculation* treatment and since we find that the contagious confidence gap is larger among individuals who exhibit base rate neglect, future work may investigate whether there are meaningful equality gains in helping individuals overcome cognitive limitations, even when these limitations are not directly related to factors such as gender. That is, insights from the broader literature on cognitive limitations and behavioral biases may prove to be particularly promising to the literature that seeks to counter discrimination and inequities.

Fourth, our results suggest caution when considering whether to remove gender information from job applications and other types of evaluations. On one hand, in light of the literature on gender-specific backlash and discrimination more generally (Riach and Rich, 2002; Rudman and Fairchild, 2004; Bowles et al., 2007; Rudman and Phelan, 2008), the removal of gender information could prove helpful. On the other hand, the removal of gender information likely decreases the chance that employers can accurately account for gender differences in confidence—even if they are provided with the training and tools to do so.

Fifth, moving beyond gender, our results highlight how—even when individuals expect some bias—they may fail to account for it. Future work may investigate whether this also proves to be the case for other biases. Future work may further explore whether expecting a bias, such as biases induced by discrimination, creates a false sense of confidence in one's ability to account for it, which may in turn hinder debiasing attempts. Indeed, as discussed in Section 5.2, we find that posterior beliefs reveal expected gender gaps in performance that are, if anything, larger for individuals who think they accurately accounted or over-accounted for the confidence gap relative to those who think that they under-accounted for it.

References

Bertrand, Marianne, G.-C. and L. F. Katz (2010): "Dynamics of the Gender Gap for Young Professionals in the Financial and Corporate Sectors," *American Economic Journal: Applied Economics*, 2.

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

Biasi, B. and H. Sarsons (Forthcoming): "Flexible Pay, Bargaining, and The Gender Gap," *Quarterly Journal of Economics*.

Blau, F. D. and L. M. Kahn (2017): "The Gender Wage Gap: Extent, Trends. and Explanations," *Journal of Economic Literature*, 55.

Bohnet, I., O. P. Hauser, and A. Kristal (2022): "Can Gender and Race Dynamics in Performance Appraisals Be Disrupted? The Case of Anchoring," .

Bohnet, I., A. van Geen, and M. Bazerman (2016): "When Performance Trumps Gender Bias: Joint vs. Separate Evaluation," *Management Science*, 62, 1225–1234.

Bohren, J. A., A. Imas, and M. Rosenberg (2019): "The Dynamics of Discrimination: Theory and Evidence," *American Economic Review*, 109, 3395–3436.

Bordalo, P., K. Coffman, N. Gennaioli, and A. Shleifer (2019): "Beliefs about Gender," *American Economic Review*, 109, 739–73.

Bowles, H. R., L. Babcock, and L. Lai (2007): "Social incentives for gender differences in the propensity to initiate negotiations: Sometimes it does hurt to ask," *Organizational Behavior and Human Decision Processes*, 103, 84–103.

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

Charness, G., R. Oprea, and S. Yuksel (2022): "How Do People Choose Between Biased Information Sources? Evidence from a Laboratory Experiment," *Journal of the European Economic Association*, 19, 1656–1691.

Coffman, K., M. Collis, and L. Kulkarni (2019a): "Stereotypes and Belief Updating," *Working Paper*.

Coffman, K., C. B. Flikkema, and O. Shurchkov (2019b): "Gender Stereotypes in Deliberation and Team Decisions," *Harvard Business School Working Paper*.

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

Coffman, K. B., M. R. Collis, and L. Kulkarni (2019c): "When to Apply?" *Working Paper*.

Coffman, K. B., C. L. Exley, and M. Niederle (2021): "The Role of Beliefs in Driving Gender Discrimination," *Management Science*, 67, 3321–3984.

Croson, R. and U. Gneezy (2009): "Gender Differences in Preferences," *Journal of Economic Literature*, 47, 448–474.

DellaVigna, S. and D. Pope (2018a): "Predicting experimental results: who knows what?" *Journal of Political Economy*, 126, 2410–2456.

——— (2018b): "What motivates effort? Evidence and expert forecasts," *The Review of Economic Studies*, 85, 1029–1069.

Enke, B. (2017): "What You See Is All There Is," *Working Paper*.

Erkal, N., L. Gangadharan, and B. H. Koh (2021): "Gender Biases in Performance Evaluation: The Role of Beliefs versus Outcomes," *SSRN Working Paper*.

Exley, C. L. and J. B. Kessler (2022): "The Gender Gap in Self-Promotion," *Quarterly Journal of Economics*, 137, 1345–1381.

Frederick, S. (2005): "Cognitive Reflection and Decision Making," *Journal of Economic Perspectives*, 19, 25–42.

Goldin, C. (2014): "A Grand Gender Convergence: Its Last Chapter," *American Economic Review*, 104, 1091–1119.

Grossman, P. J., C. Eckel, M. Komai, and W. Zhan (2019): "It pays to be a man: Rewards for leaders in a coordination gam," *Journal of Economic Behavior & Organization*, 161, 197–215.

Hernandez-Arenaz, I. and N. Iriberri (2019): "A review of gender differences in negotiation," *Oxford Research Encyclopedia of Economics and Finance*.

Huber, C. and J. Huber (2020): "Bad bankers no more? Truth-telling and (dis) honesty in the finance industry," *Journal of Economic Behavior & Organization*, 180, 472–493.

Kahneman, D. and A. Tversky (1972): "On Prediction and Judgment," *ORI Research Monograph*, 12.

Kolev, J., Y. Fuentes-Medel, and F. Murray (2019): "Is Blinded Review Enough? How Gendered Outcomes Arise Even Under Anonymous Evaluation," *Academy of Management Proceedings*, 1.

Lundeberg, M. A., P. W. Fox, and J. Punćcohaŕ (1994): "Highly confident but wrong: Gender differences and similarities in confidence judgments." *Journal of educational psychology*, 86.

Michelmore, K. and S. Sassler (2016): "Explaining the gender wage gap in STEM: Does field sex composition matter?" *RSF: The Russell Sage Foundation Journal of the Social Sciences*, 2, 194–215.

Mobius, M. M., M. Niederle, P. Niehaus, and T. S. Rosenblat (2022): "Managing Self-Confidence: Theory and Experimental Evidence," *Management Science*.

Murciano-Goroff, R. (2021): "Missing Women in Tech: The Labor Market for Highly Skilled Software Engineers," *Management Science*.

Niederle, M. (2016): "Gender," in *Handbook of Experimental Economics*, ed. by J. Kagel and A. E. Roth, Princeton University Press, vol. 2, 481–553.

Niederle, M. and L. Vesterlund (2007): "Do Women shy away from competition? Do men compete too much?" *Quarterly Journal of Economics*, 122, 1067–1101.

——— (2011): "Gender and Competition," *Annual Review of Economics*, 3, 601–630.

Reuben, E., P. Sapienza, and L. Zingales (2014): "How stereotypes impair women's careers in science," *Proceedings of the National Academy of Sciences*, 111, 4403–4408.

Reuben, Ernesto, R.-B. P. S. P. and L. Zingales (2012): "The Emergence of Male Leadership in Competitive Environments," *Journal of Economic Behavior & Organization*, 83.

Reuben, Ernesto, W.-M. and B. Zafar (2017): "Preferences and Biases in Educational Choices and Labor Market Expectations: Shrinking the Black Box of Gender," *The Economic Journal*, 627, 604.

Riach, P. A. and J. Rich (2002): "Field Experiments of Discrimination in the Market Place," *The Economic Journal*, 112.

Roussille, N. (2021): "The central role of the ask gap in gender pay inequality," *Working Paper*.

Rudman, L. A. and K. Fairchild (2004): "Reactions to Counterstereotypic Behavior: The Role of Backlash in Cultural Stereotype Maintenance," *Journal of Personality and Social Psychology*, 87.

Rudman, L. A. and J. E. Phelan (2008): "Backlash effects for disconfirming gender stereotypes in organizations," *Research in organizational behavior*, 28.

Saccardo, S. and M. Serra-Garcia (2022): "Enabling or Limiting Cognitive Flexibility? Evidence of Demand for Moral Commitment," *Working Paper*.

Sarsons, H. and X. Guo (2021): "Confidence Men? Evidence on Confidence and Gender among Top Economists," *American Economic Association Papers and Proceedings*, 111.

Appendix Table of Contents

A Additional Design Details

Figure A.1: Screenshot of Classifier Question in the *Worker Study*

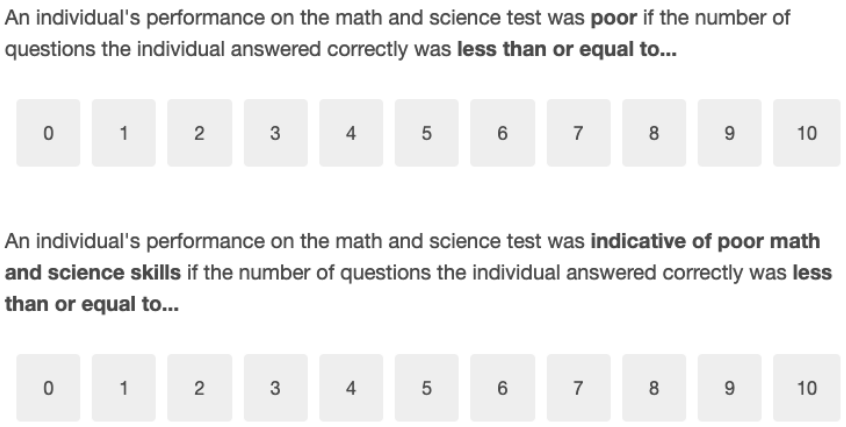


Figure A.2: Screenshot of Main Self-Evaluation in the *Worker Study*

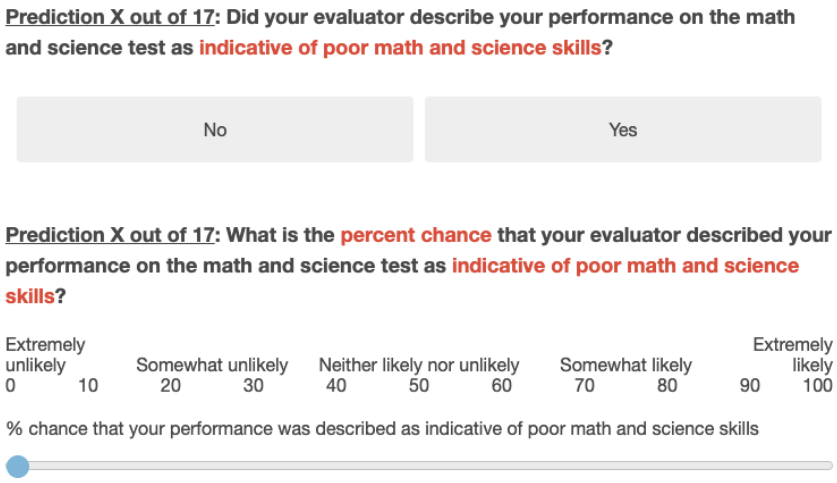


Figure A.3: Screenshot of Posterior Belief in *Baseline* and *Attention* Treatment of the *Evaluator Study*

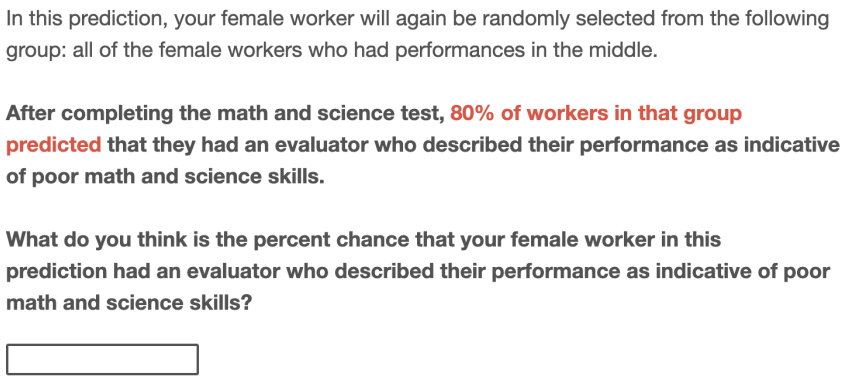


Figure A.4: Screenshot of Posterior Belief in the *Calculation* Treatment of the *Evaluator Study*

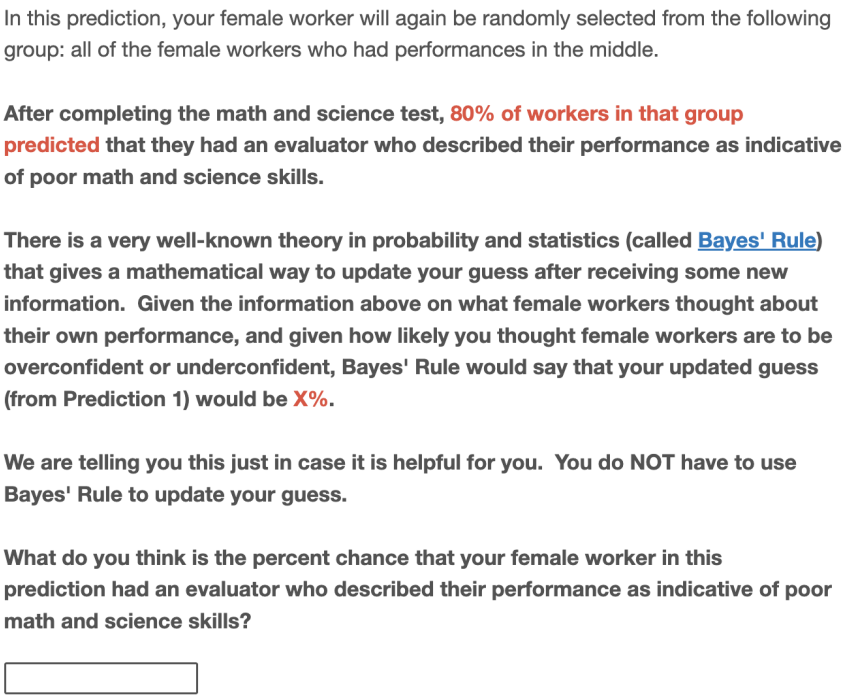
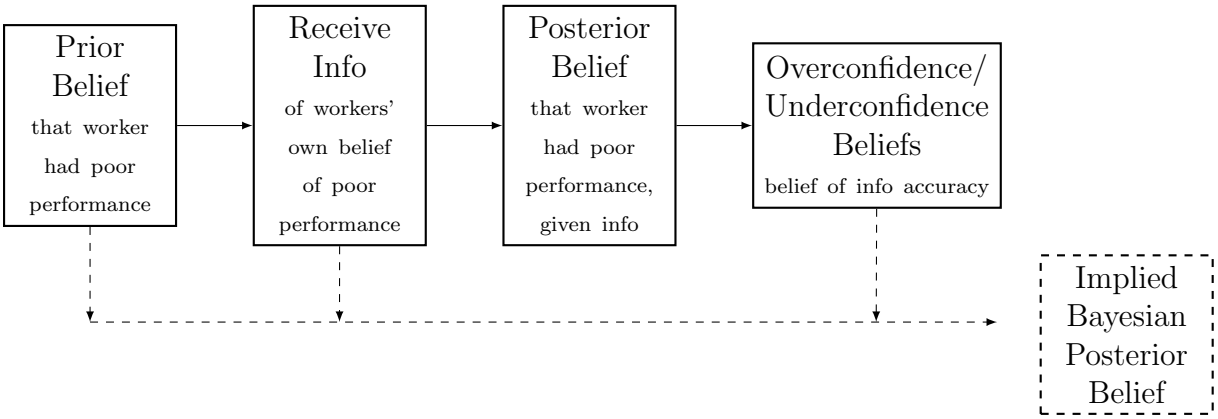
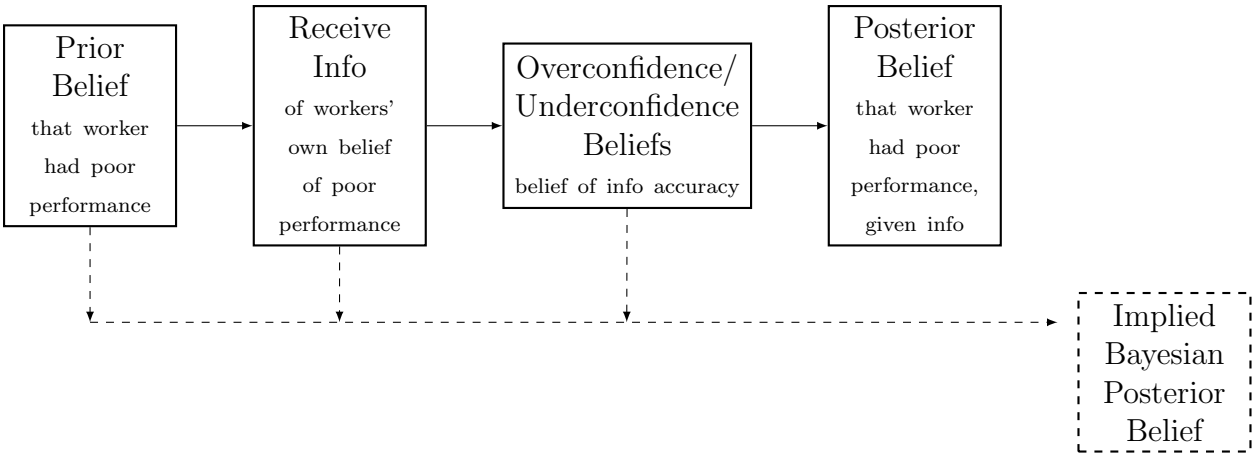


Figure A.5: Timeline of *Baseline* and *Baseline, Unknown Gender* treatments of the *Evaluator Study*



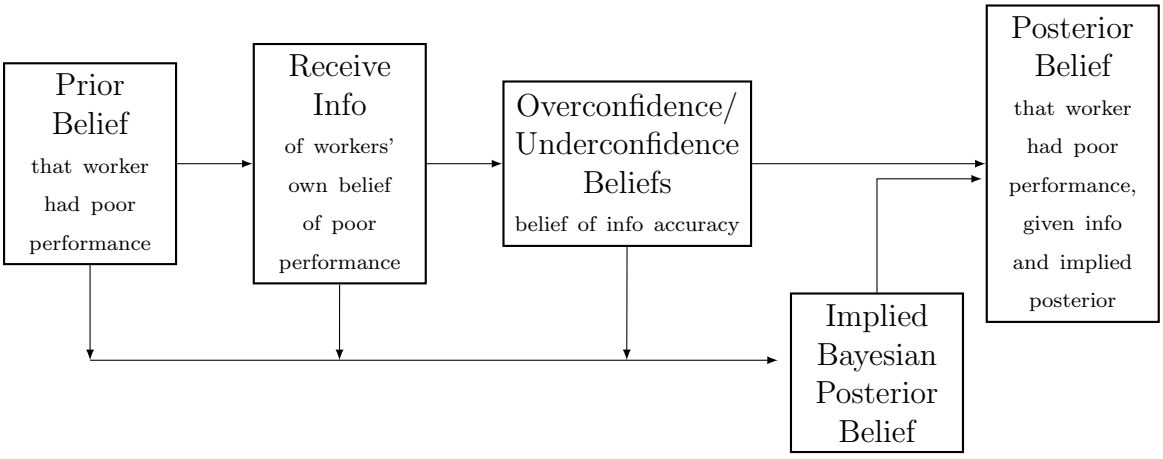
In the *Baseline* and *Baseline, Unknown Gender* treatments, we elicit an evaluator's prior belief that a randomly selected male or female worker had a poor performance. Then, we provide evaluators with the percentage of male or female workers who believed they had a poor performance. After this, we elicit posterior beliefs that a randomly selected male or female worker had a poor performance. Finally, we elicit evaluators' beliefs of the percentage of male or female workers they believe to be overconfident and underconfident conditional on actual performance. The prior beliefs, signal, and over/underconfidence beliefs combine to form the implied Bayesian posterior belief, but evaluators never see this implied belief.

Figure A.6: Timeline of *Attention* and *Attention, Unknown Gender* treatments of the *Evaluator Study*



In the *Attention* and *Attention, Unknown Gender* treatments, we elicit an evaluator's prior belief that a randomly selected male or female worker had a poor performance. Then, we provide evaluators with the percentage of male or female workers who believed they had a poor performance. After this, we elicit evaluators' beliefs of the percentage of male or female workers they believe to be overconfident and underconfident conditional on actual performance. Finally, we elicit posterior beliefs that a randomly selected male or female worker had a poor performance. The prior beliefs, signal, and over/underconfidence beliefs combine to form the implied Bayesian posterior belief, but evaluators never see this implied belief.

Figure A.7: Timeline of *Calculation* and *Calculation, Unknown Gender* treatments of the *Evaluator Study*



In the *Calculation* and *Calculation, Unknown Gender* treatments, we elicit an evaluator's prior belief that a randomly selected male or female worker had a poor performance. Then, we provide evaluators with the percentage of male or female workers who believed they had a poor performance. After this, we elicit evaluators' beliefs of the percentage of male or female workers they believe to be overconfident and underconfident conditional on actual performance. The prior beliefs, signal, and over/underconfidence beliefs combine to form the implied Bayesian posterior belief. We show this implied Bayesian posterior belief to subjects in the final part of the study when we elicit posterior beliefs that a randomly selected male or female worker had a poor performance.

Table A.1: Questions in the *Worker* Study

Q#	Question Text	Answer
CQ1	An individual's performance on the math and science test was indicative of poor math and science skills if the number of questions the individual answered correctly was less than or equal to ____.	0–10
CQ2	An individual's performance on the math and science test was poor if the number of questions the individual answered correctly was less than or equal to ____.	0–10
0	Out of the 10 questions on the math and science test, what do you think is the number you answered correctly?	0–10
1B	Did you get 3 or more questions right out of the 10 questions on the math and science test?	yes or no
1C	What is the percent chance that you got 3 or more questions right out of the 10 questions on the math and science test?	0%–100%
2B	Did you get 5 or more questions right out of the 10 questions on the math and science test?	yes or no
2C	What is the percent chance that you got 5 or more questions right out of the 10 questions on the math and science test?	0%–100%
3B	Did you get 7 or more questions right out of the 10 questions on the math and science test?	yes or no
3C	What is the percent chance that you got 7 or more questions right out of the 10 questions on the math and science test?	0%–100%
4B	Did you score in the top half when compared to other participants who took the study?	yes or no
4C	What is the percent chance that you scored in the top half when compared to other participants who took the study?	0%–100%
5B	Did you score in the top half when compared to women who took the study?	yes or no
5C	What is the percent chance that you scored in the top half when compared to women who took the study?	0%–100%
6B	Did you score in the top half when compared to men who took the study?	yes or no
6C	What is the percent chance that you scored in the top half when compared to men who took the study?	0%–100%
7B	Did your evaluator describe your performance on the math and science test as poor?	yes or no
7C	What is the percent chance that your evaluator described your performance on the math and science test as poor?	0%–100%
8B	Did your evaluator describe your performance on the math and science test as indicative of poor math and science skills?	yes or no
8C	What is the percent chance that your evaluator described your performance on the math and science test as indicative of poor math and science skills?	0%–100%

CC1 and CC2, the two classifier questions, appeared together on the same page before the instructions for the self-evaluations. Self-Evaluation 0 appears on its own decision screen, and all other self-evaluations appears in pairs on a decision screen. Specifically, on a decision screen, the first question is Self-Evaluation iB and the second question is Self-Evaluation iC for $i = 1, 2, .., 8$. The order of the resulting 9 decision screens is randomized at the worker level. Self-Evaluation 0 involves an integer guess from 0-10, and they earn \$1 in that self-evaluation if their guess is correct. Self-Evaluations iB (for $i = 1, 2, .., 8$) involve a binary guess (yes/no), and they earn \$1 in each of those self-evaluations if their guess is correct. Self-Evaluations iC (for $i = 1, 2, .., 8$) ask them to guess a percent chance of some outcome being true (0-100%), and they earn a \$1 bonus in each of those self-evaluations according to an incentive-compatible BDM procedure.

Table A.2: Beliefs in the *Evaluator* Study

Q Label	Question Text
Prior Belief	What do you think is the percent chance that your male/female worker in this prediction had a classifier who described their performance as indicative of poor math and science skills?
Posterior Belief	After completing the math and science test, 56%/80% of male/female workers predicted that their classifier described their performance as indicative of poor math and science skills. What do you think is the percent chance that your male/female worker in this prediction had a classifier who described their performance as indicative of poor math and science skills?
Overconfidence Belief	If your male/female worker in this prediction had a classifier who described their performance as indicative of poor math and science skills, what do you think is the percent chance that your male/female worker is overconfident because they predicted that their classifier did NOT describe their performance as indicative of poor math and science skills?
Underconfidence Belief	If your male/female worker in this prediction had a classifier who did NOT describe their performance as indicative of poor math and science skills, what do you think is the percent chance that your male/female worker is underconfident because they predicted that their classifier described their performance as indicative of poor math and science skills?

The above table describes the exact wording of the belief questions—with the exception of "evaluator" being replaced with "classifier" as explained in footnote 8—elicited in the *Evaluator* Study for the treatments in which the gender of the workers is known (and note that each evaluator is only asked about male workers or only asked about female workers). For the treatments in which the gender of the worker is unknown, male/female is replaced with group-1/group-2. Also, recall that—as described in Section 2—we define a worker as having a "poor performance" if their classifier indicated their performance was indicative of poor math and science skills in response to Classifier Question 1 (CC1 in Appendix Table A.1), and then use the "poor performance" shorthand throughout our main text. Each belief question asks evaluators to guess a percent chance of some outcome being true (0-100%), and they earn a $1 bonus in each of those self-evaluations according to an incentive-compatible BDM procedure. The overconfidence belief and underconfidence belief are always shown on the same decision screen. All other beliefs are shown on separate decision screens. In *Baseline* and *Baseline, Unknown Gender* treatments, we elicit prior beliefs, then posterior beliefs, and then over/underconfidence beliefs. In the *Attention* and *Calculation* treatments (for both known and unknown gender), we elicit over/underconfidence beliefs before posterior beliefs.

Table A.3: Beliefs in the *Evaluator (Additional Questions)* Study

Q Label	Question Text
Prior (3+)	What do you think is the percent chance that your male/female worker in this prediction got 3 or more questions right?
Prior (5+)	Same as Prior (3+) but replace 3 with 5
Prior (7+)	Same as Prior (3+) but replace 3 with 7
Prior (poor-2)	What do you think is the percent chance that your male/female worker in this prediction had a classifier who described his/her performance as poor?
Prior (top half)	What do you think is the percent chance that your male/female worker in this prediction scored in the top half?
Posterior (3+)	After completing the math and science test, AVG% of male/female workers predicted that they got 3 or more questions right. What do you think is the percent chance that your male/female worker in this prediction got 3 or more questions right?
Posterior (5+)	Same as Posterior (3+) but replace 3 with 5
Posterior (7+)	Same as Posterior (3+) but replace 3 with 7
Posterior (poor-2)	After completing the math and science test, AVG% of male/female workers predicted that they had a classifier who described their performance as poor. What do you think is the percent chance that your male/female worker in this prediction had a classifier who described his/her performance as poor?
Posterior (top half)	After completing the math and science test, AVG% of male/female workers predicted that they scored in the top half. What do you think is the percent chance that your male/female worker in this prediction scored in the top half?
Overconfidence (3+)	If your male/female worker in this prediction got fewer than 3 questions right, what do you think is the percent chance that your male/female worker is overconfident because they predicted that they got 3 or more questions right?
Overconfidence (5+)	Same as Overconfidence (3+) but replace 3 with 5
Overconfidence (7+)	Same as Overconfidence (3+) but replace 3 with 7
Overconfidence (poor-2)	If your male/female worker in this prediction had a classifier who described his/her performance as poor, what do you think is the percent chance that your male/female worker is overconfident because they predicted that their classifier did not describe their performance as poor?
Overconfidence (top half)	If your male/female worker in this prediction did not score in the top half, what do you think is the percent chance that your male/female worker is overconfident because they predicted that scored in the top half?
Underconfidence (3+)	If your male/female worker in this prediction got more than 3 questions right, what do you think is the percent chance that your male/female worker is underconfident because they predicted that they got fewer than 3 questions right?
Underconfidence (5+)	Same as Underconfidence (3+) but replace 3 with 5
Underconfidence (7+)	Same as Underconfidence (3+) but replace 3 with 7
Underconfidence (poor-2)	If your male/female worker in this prediction had a classifier who did not describe his/her performance as poor, what do you think is the percent chance that your male/female worker is underconfident because they predicted that their classifier described their performance as poor?
Underconfidence (top galf)	If your male/female worker in this prediction scored in the top half, what do you think is the percent chance that your male/female worker is underconfident because they predicted that did not score in the top half?

This table describes the exact wording of the additional belief questions—with the exception of "evaluator" being replaced with "classifier" as explained in footnote 8—elicited in the *Evaluator (Alternative Questions)* Study. Each belief question asks evaluators to guess a percent chance of some outcome being true (0-100%), and they earn a $1 bonus in each of those self-evaluations according to an incentive-compatible BDM procedure. The overconfidence and underconfidence belief are always shown on the same decision screen. All other beliefs are shown on separate decision screens. We elicit the block of 6 prior beliefs, then the block of 6 posterior beliefs, and then the block of 12 over/underconfidence beliefs. The order of the beliefs within each block is randomized.

Table A.4: Overview of The Evaluator Study Treatment Conditions

Study Version	Description	Sample Size, Date	Paper Section
Evaluator Study – Baseline Treatment	Elicit prior belief, posterior belief, overconfidence and underconfidence beliefs (in that order) about main self-evaluation question, randomized to provide beliefs about either male or female workers	N=402, July 2022	Section 4.1
Evaluator Study – Attention Treatment	Same as Baseline Treatment except overconfidence and underconfidence beliefs elicited before posterior belief	N=403, July 2022	Section 4.2
Evaluator Study – Calculation Treatment	Same as Attention Treatment except provided with implied Bayesian posterior while reporting posterior beliefs	N=405, July 2022	Section 4.2
Evaluator Study – Baseline, Unknown Gender Treatment	Same as Baseline Treatment except the gender of workers is unknown	N=405, July 2022	Section 4.3
Evaluator Study – Attention, Unknown Gender Treatment	Same as Attention Treatment except the gender of workers is unknown	N=392, July 2022	Section 4.3
Evaluator Study – Calculation, Unknown Gender Treatment	Same as Calculation Treatment except the gender of workers is unknown	N=393, July 2022	Section 4.3

This table provides a brief overview of the 6 treatments run as part of the *Evaluator Study*. Evaluators were randomized into one of these 6 treatments. Evaluators were further randomized to evaluate either male or female workers.

Table A.5: Overview of Additional Evaluator Study Versions

Study Version	Description	Sample Size, Date	Paper Section
Evaluator (Professional Evaluators) Study – Baseline Treatment	Same as Evaluator Study – Baseline Treatment except that we recruit evaluators who have experience making hiring experience and in management, and workers are from the Worker (Undergraduates) Study	N=409, September 2022	Section 6.1
Evaluator (Professional Evaluators) Study – Baseline, Unknown Gender Treatment	Same as the Baseline Treatment except the gender of workers is unknown	N=391, September 2022	Section 6.1
Evaluator (Extended) Study – Baseline Treatment	Same as Evaluator Study – Baseline Treatment except that, before providing posterior belief, evaluators make 20 predictions about specific workers after learning each of those workers' self-evaluations	N=406, May 2022	Sections 6.2 6.3
Evaluator (Extended) Study – Strategic Incentives Treatment	Same as Evaluator (Extended) Study – Baseline Treatment except that they provide beliefs about workers who, rather facing accuracy incentives, faced strategic incentives to inflate self-evaluations	N=394, May 2022	Section 6.4
Evaluator (Extended) Study – Joint Evaluations Treatment	Same as Evaluator (Extended) Study – Baseline Treatment except that, rather than providing beliefs only about men or women, they simultaneously provide beliefs about men and women	N=205, May 2022	Section 6.5
Evaluator (Extended) Study – Joint Evaluations, Strategic Incentives Treatment	Same as Evaluator (Extended) Study – Joint Evaluations Treatment except that they provide beliefs about workers who faced strategic incentives to inflate self-evaluations (rather than workers who are incentivized to accurately report self-evaluations)	N=195, May 2022	Section 6.5
Evaluator (Alternative Questions) Study	Same as Evaluator Study – Baseline Treatment except that, rather than only answering the belief questions in Appendix Table A.2, also answering the belief questions in Appendix Table A.3	N=400, May 2022	Section 6.7
Evaluator (Additional Demographics) Study	Same as Evaluator Study – Baseline Treatment except that, rather than providing beliefs about men or women, they provide beliefs about men or women who who work full time, are between 26 and 40 years old, live in the Southern region of the United States, and have completed at least some college education	N=198, May 2022	Section 6.8
Evaluator (Known Performance) Study	Same as Evaluator Study – Baseline Treatment except that, rather than only providing beliefs about men and women, asked to provide beliefs about men who got 5 questions right on the test or women who got 5 questions right on the test	N=198, May 2022	Section 6.9

This table provides a brief overview of the additional study versions we ran. Evaluators in the Evaluator (Extended) Study were randomized into one of the 4 treatments described above.

Table A.6: Overview of The Worker Study Versions

Study Version	Description	Sample Size, Date	Paper Section
Worker Study – Baseline Treatment	10-question math and science test followed by 17 self-evaluations shown in Appendix Table A.1	N=393, April 2022	Section 3
Worker Study – Strategic Incentives	Same the Baseline Treatment but workers faced strategic incentives to inflate self-evaluations	N=387, April 2022	Section 6.4
Worker (Undergraduates) Study	Workers were Ohio State University undergraduates who completed a 10-question math and science test followed by 13 self-evaluations. Rather than earning 10 cents for each question they answer correctly on the math and science test in Part 1, they earn $1 for each question they answer correctly. Rather than having a chance of earning $1 for each prediction they make in Part 1, they have a chance of earning $10 for each prediction they make in Part 1. Furthermore, some of the easiest questions in the Worker Study are replaced with more difficult questions in the Worker (Undergraduates) Study. Finally, workers in this study answered the questions in Appendix Table A.1 except for questions 4B, 4C, 5B, 5C, 6B, and 6C. In addition to these questions, workers answered Question 9B: "Did you get 9 or more questions right out of the 10 questions on the math and science test?" and Question 9C: "What is the percent chance that you got 9 or more questions right out of the 10 questions on the math and science test?"	N=350, March/April 2022	Section 6.1

This table provides a brief overview of the 3 worker study versions. Workers recruited for the first 2 study versions were randomized into one of them.

B Additional Results

Table B.1: Self-Evaluations in the *Worker Study*

Panel A: Self-Evaluations about Absolute Performance (Q# = 0-3C)

	0	1B	1C	2B	2C	3B	3C
Female	-0.54***	-0.09**	-9.40***	-0.11***	-5.68**	-0.05*	-3.30
	(0.16)	(0.04)	(2.66)	(0.04)	(2.69)	(0.03)	(2.58)
N	393	393	393	393	393	393	393
Perf FE	yes	yes	yes	yes	yes	yes	yes

Panel B: Self-Evaluations (Q# 4B-6C) about Relative Performance

	4B	4C	5B	5C	6B	6C
Female	-0.11**	-7.15***	-0.08*	-7.39***	-0.13***	-9.11***
	(0.04)	(2.59)	(0.05)	(2.52)	(0.05)	(2.58)
N	393	393	393	393	393	393
Perf FE	yes	yes	yes	yes	yes	yes

Panel C: Self-Evaluations (Q# 7B-8C) about Subjective Performance

	7B	7C	8B	8C
Female	0.14***	10.64***	0.16***	7.79***
	(0.04)	(2.49)	(0.04)	(2.59)
N	393	393	393	393
Perf FE	yes	yes	yes	yes

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are robust. Results are from OLS regressions of the responses provided to the self-evaluation question noted in each column (see Appendix Table A.1 for details on each self-evaluation question). The responses to the binary self-evaluation questions are coded as 1 if the worker answers "yes" or 0 if the worker answers "no." *Female* is an indicator for the worker identifying as a woman. Perf FEs are dummies for each possible performance out of the 10 questions on the test. Data are from the 393 participants who identified as a man or a woman in the *Worker Study*.

Table B.2: Evaluators' Beliefs in the *Attention* treatment of the *Evaluator Study*

	Prior	Over-confidence	Under-confidence	Implied Bayesian Posterior	Posterior
	(1)	(2)	(3)	(4)	(5)
Panel A: Evaluators' Beliefs					
B(F)	42.41	42.69	52.77	43.69	58.92
B(M)	39.00	47.30	42.93	40.15	48.07
Δ	3.41*	-4.60**	9.84***	3.54*	10.85***
	(1.83)	(2.20)	(2.08)	(1.80)	(1.73)
Panel B: Evaluators' Beliefs - Truth					
B(F) - Truth(F)	-7.121	27.34	-22.03	-5.840	9.386
B(M) - Truth(M)	-8.795	8.235	-9.210	-7.640	0.280
Δ - Truth(Δ)	1.67	19.11***	-12.82***	1.80	9.11***
	(1.83)	(2.20)	(2.08)	(1.80)	(1.73)
N	403	403	403	403	403
Truth(F)	49.53	15.35	74.80	49.53	49.53
Truth(M)	47.79	39.06	52.14	47.79	47.79
Truth(Δ)	1.74	-23.70	22.65	1.74	1.74

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are robust and shown in parentheses. Results follow the structure of Table 2. Data are from the 403 participants in the *Attention* treatment of *Evaluator Study*.

Table B.3: Evaluators' Beliefs in the *Calculation* treatment of the *Evaluator Study*

	Prior	Over-confidence	Under-confidence	Implied Bayesian Posterior	Posterior
	(1)	(2)	(3)	(4)	(5)
Panel A: Evaluators' Beliefs					
B(F)	41.72	39.70	55.06	42.48	48.06
B(M)	38.65	49.12	43.33	39.37	43.15
Δ	3.07*	-9.42***	11.73***	3.11*	4.92***
	(1.82)	(2.27)	(1.98)	(1.75)	(1.81)
Panel B: Evaluators' Beliefs - Truth					
B(F) -Truth	-7.812	24.35	-19.74	-7.055	-1.466
B(M) - Truth	-9.145	10.06	-8.805	-8.424	-4.642
Δ - Truth(Δ)	1.33	14.29***	-10.93***	1.37	3.18*
	(1.82)	(2.27)	(1.98)	(1.75)	(1.81)
N	405	405	405	405	405
Truth(F)	49.53	15.35	74.80	49.53	49.53
Truth(M)	47.79	39.06	52.14	47.79	47.79
Truth(Δ)	1.74	-23.70	22.65	1.74	1.74

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are robust and shown in parentheses. Results follow the structure of Table 2. Data are from the 405 participants in the *Calculation* treatment of *Evaluator Study*.

Table B.4: Evaluators' Beliefs in the *Baseline, Unknown Gender* treatment of the *Evaluator Study*

	Prior	Over-confidence	Under-confidence	Implied Bayesian Posterior	Posterior
	(1)	(2)	(3)	(4)	(5)
Panel A: Evaluators' Beliefs					
B(F)	38.39	37.70	48.29	42.12	61.65
B(M)	40.53	40.72	45.13	41.83	50.59
Δ	-2.14	-3.02	3.16	0.29	11.06***
	(1.74)	(2.15)	(2.08)	(1.73)	(1.61)
Panel B: Evaluators' Beliefs - Truth					
B(F) - Truth(F)	-11.14	22.35	-26.51	-7.41	12.12
B(M) - Truth(M)	-7.26	1.66	-7.01	-5.96	2.80
Δ - Truth(Δ)	-3.88**	20.69***	-19.50***	-1.45	9.32***
	(1.74)	(2.15)	(2.08)	(1.73)	(1.61)
N	405	405	405	405	405
Truth(F)	49.53	15.35	74.80	49.53	49.53
Truth(M)	47.79	39.06	52.14	47.79	47.79
Truth(Δ)	1.74	-23.70	22.65	1.74	1.74

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are robust and shown in parentheses. Results follow the structure of Table 2. Data are from the 405 participants in the *Baseline, Unknown Gender* treatment of *Evaluator Study*.

Table B.5: Evaluators' Beliefs in the *Attention, Unknown Gender* treatment of the *Evaluator Study*

	Prior	Over-confidence	Under-confidence	Implied Bayesian Posterior	Posterior
	(1)	(2)	(3)	(4)	(5)
Panel A: Evaluators' Beliefs					
B(F)	40.71	40.39	52.74	42.37	59.09
B(M)	39.43	46.90	45.69	40.02	48.53
Δ	1.28	-6.50***	7.06***	2.35	10.56***
	(1.95)	(2.35)	(2.10)	(1.89)	(1.74)
Panel B: Evaluators' Beliefs - Truth					
B(F) - Truth(F)	-8.82	25.04	-22.06	-7.16	9.56
B(M) - Truth(M)	-8.36	7.84	-6.45	-7.77	0.74
Δ - Truth(Δ)	-0.46	17.21***	-15.60***	0.61	8.82***
	(1.95)	(2.35)	(2.10)	(1.89)	(1.74)
N	392	392	392	388	392
Truth(F)	49.53	15.35	74.80	49.53	49.53
Truth(M)	47.79	39.06	52.14	47.79	47.79
Truth(Δ)	1.74	-23.70	22.65	1.74	1.74

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are robust and shown in parentheses. Results follow the structure of Table 2. Data are from the 392 participants in the *Attention, Unknown Gender* treatment of *Evaluator Study*. Sample size differs slightly in column (4) as some evaluators' beliefs imply a Bayesian posterior that is undefined.

Table B.6: Evaluators' Beliefs in the *Calculation, Unknown Gender* treatment of the *Evaluator Study*

	Prior	Over-confidence	Under-confidence	Implied Bayesian Posterior	Posterior
	(1)	(2)	(3)	(4)	(5)
Panel A: Evaluators' Beliefs					
B(F)	41.23	38.39	50.03	44.36	49.07
B(M)	40.62	46.02	47.02	40.84	44.20
Δ	0.61	-7.63***	3.01	3.53**	4.87***
	(1.82)	(2.24)	(2.12)	(1.76)	(1.77)
Panel B: Evaluators' Beliefs - Truth					
B(F) - Truth(F)	-8.30	23.04	-24.77	-5.17	-0.46
B(M) - Truth(M)	-7.17	6.96	-5.12	-6.95	-3.59
Δ - Truth(Δ)	-1.13	16.08***	-19.65***	1.79	3.13*
	(1.82)	(2.24)	(2.12)	(1.76)	(1.77)
N	393	393	393	392	393
Truth(F)	49.53	15.35	74.80	49.53	49.53
Truth(M)	47.79	39.06	52.14	47.79	47.79
Truth(Δ)	1.74	-23.70	22.65	1.74	1.74

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are robust and shown in parentheses. Results follow the structure of Table 2. Data are from the 393 participants in the *Calculation, Unknown Gender* treatment of *Evaluator Study*. Sample size differs slightly in column (4) as some evaluators' beliefs imply a Bayesian posterior that is undefined.

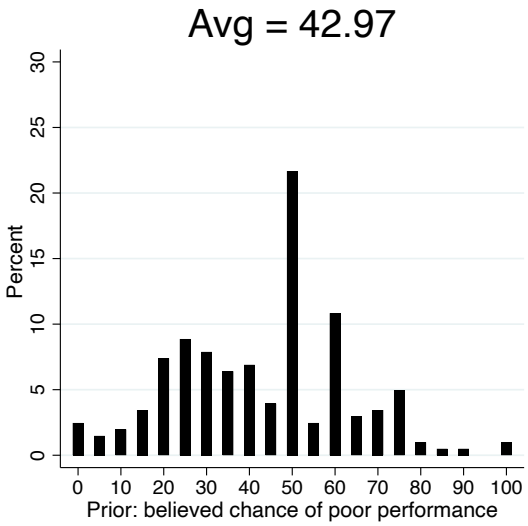
Table B.7: Evaluators' Beliefs in the *Baseline, Unknown Gender*, *Attention, Unknown Gender* and *Calculation, Unknown Gender* treatment of the *Evaluator Study*

	Prior	Over-confidence	Under-confidence	Implied Bayesian Posterior	Posterior
	(1)	(2)	(3)	(4)	(5)
Panel A: Evaluators' Beliefs					
Δ	-2.14	-3.02	3.16	0.29	11.06***
	(1.74)	(2.15)	(2.08)	(1.73)	(1.61)
Δ*Attention	3.43	-3.49	3.89	2.06	-0.50
	(2.62)	(3.18)	(2.96)	(2.56)	(2.37)
Δ*Calculation	2.76	-4.61	-0.15	3.24	-6.19***
	(2.52)	(3.11)	(2.97)	(2.47)	(2.39)
Panel B: Evaluators' Beliefs - Truth					
Δ	-3.88**	20.69***	-19.50***	-1.45	9.32***
	(1.74)	(2.15)	(2.08)	(1.73)	(1.61)
Δ*Attention	3.43	-3.49	3.89	2.06	-0.50
	(2.62)	(3.18)	(2.96)	(2.56)	(2.37)
Δ*Calculation	2.76	-4.61	-0.15	3.24	-6.19***
	(2.52)	(3.11)	(2.97)	(2.47)	(2.39)
N	1190	1190	1190	1185	1190
Condition FE	yes	yes	yes	yes	yes
Truth(Δ)	1.74	-23.70	22.65	1.74	1.74

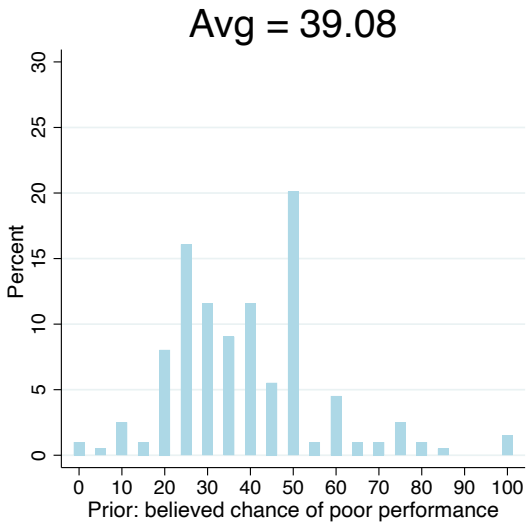
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are robust and shown in parentheses. Results follow the structure of Table 3. Data are from the 1190 participants in the *Baseline, Unknown Gender* treatment, the *Attention, Unknown Gender* or the *Calculation, Unknown Gender* treatment of *Evaluator Study*. Sample size differs slightly in column (4) as some evaluators' beliefs imply a Bayesian posterior that is undefined.

Figure B.1: *Baseline Treatment*: Prior and Posterior Beliefs

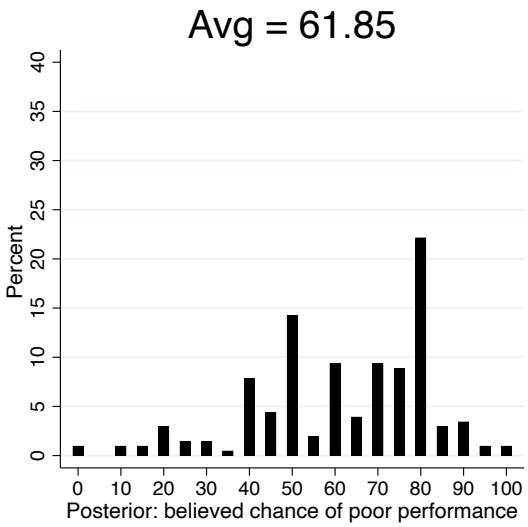
A: Prior Beliefs about Women



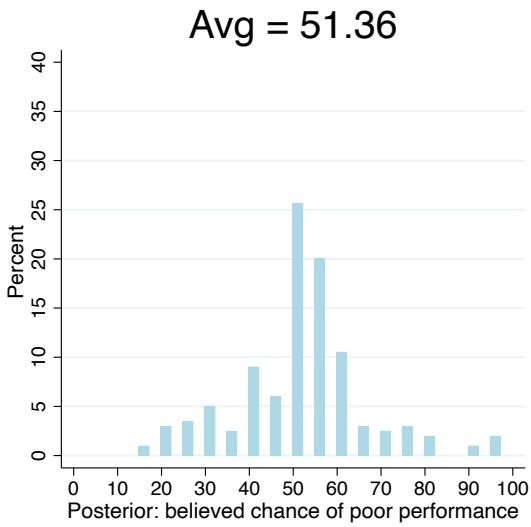
B: Prior Beliefs about Men



C: Posterior Beliefs about Women



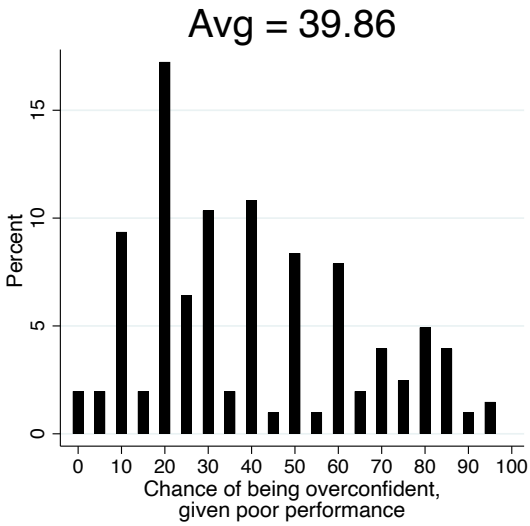
D: Posterior Beliefs about Men



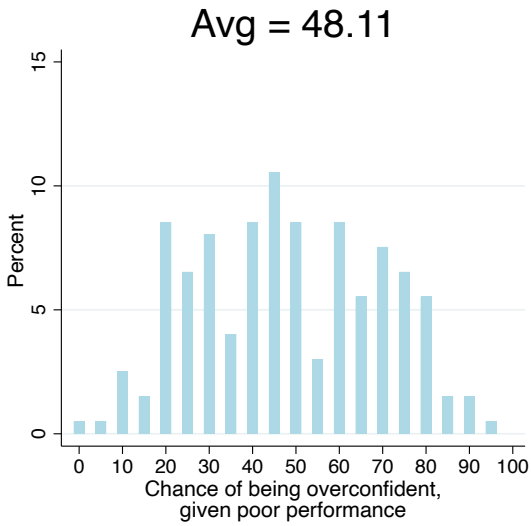
Data are from *Baseline* treatment of the *Evaluator Study*.

Figure B.2: *Baseline Treatment*: Confidence Beliefs

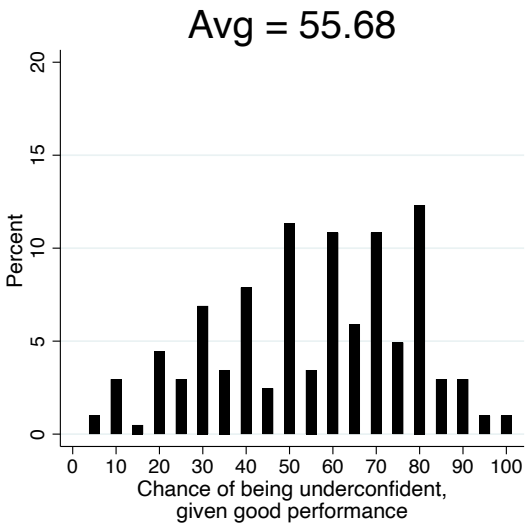
A: Overconfidence Beliefs about Women



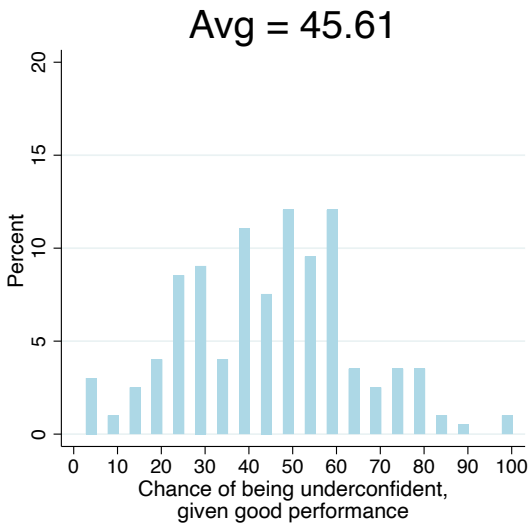
B: Overconfidence Beliefs about Men



C: Underconfidence Beliefs about Women



D: Underconfidence Beliefs about Men

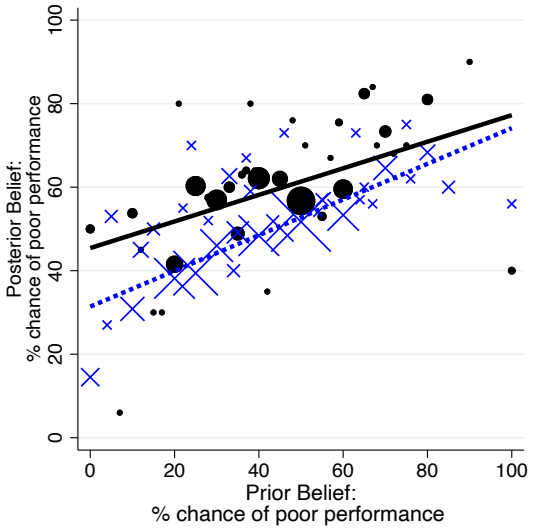


Data are from *Baseline* treatment of the *Evaluator Study*.

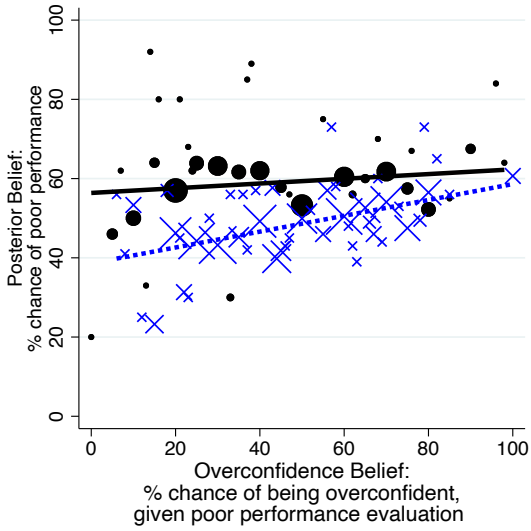
C Additional Heterogeneity Results

Figure C.1: *Attention Treatment*: Posterior Beliefs as a Function of Their Other Beliefs

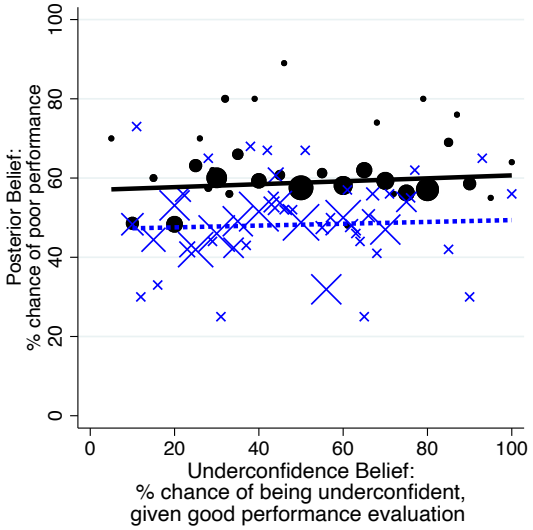
A: As a function of their Prior Beliefs



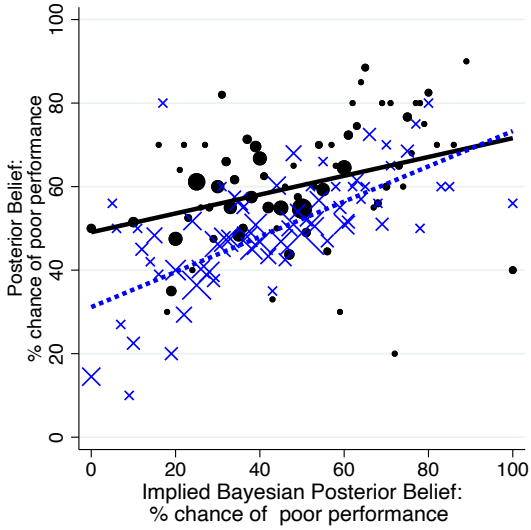
B: As a function of their Overconfidence Beliefs



C: As a function of their Underconfidence Beliefs

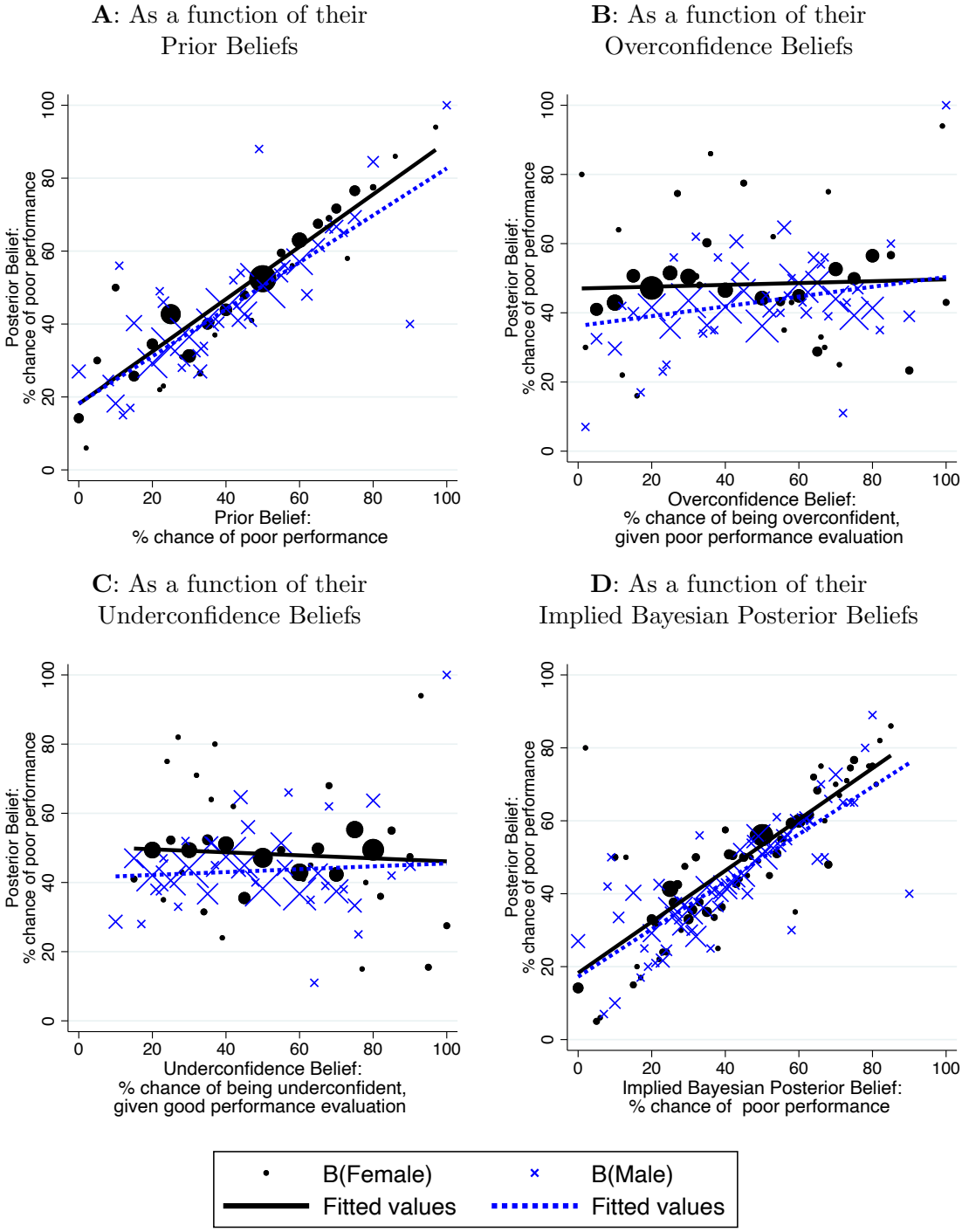


D: As a function of their Implied Bayesian Posterior Beliefs



See Figure 2 for a description of the graphs above. Data are from *Attention* treatment of the *Evaluator Study*.

Figure C.2: *Calculation Treatment*: Posterior Beliefs as a Function of Their Other Beliefs



See Figure 2 for a description of the graphs above. Data are from the *Calculation* treatment of the *Evaluator Study*.

Table C.1: By believed gender differences in confidence: evaluators' posterior beliefs about workers in *Evaluator Study* when gender is known

	Gender difference in confidence:			Gender difference in confidence in STEM:		
	Women less confident	No difference	Women more confident	Women less confident	No difference	Women more confident
	(1)	(2)	(3)	(4)	(5)	(6)
Δ	10.96***	9.91***	12.83	15.01***	8.98***	-16.40*
	(2.48)	(2.57)	(13.52)	(2.19)	(2.86)	(8.42)
Δ*Attention	0.61	0.03	-3.67	-1.66	-0.43	22.02*
	(3.45)	(3.68)	(18.27)	(3.22)	(4.04)	(11.39)
Δ*Calculation	-3.81	-7.06*	-13.01	-6.69**	-6.34	10.36
	(3.52)	(3.70)	(17.08)	(3.26)	(4.05)	(10.71)
N	621	555	34	622	508	80
Condition FE	yes	yes	yes	yes	yes	yes
Truth(Δ)	1.74	1.74	1.74	1.74	1.74	1.74

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are robust and shown in parentheses. The data are from the *Baseline*, *Attention*, and *Calculation* treatments for the group of evaluators noted in the column, specifically evaluators who, in the follow-up survey, indicate that they believe that: women are less confident than men in Column 1, there is no gender differences in confidence in Column 2, women are more confident than men in Column 3, women are less confident than men in STEM fields in Column 4, there is no gender differences in confidence in STEM in Column 5, and women are more confident than men in STEM fields in Column 6. The regression specifications are the same as in Appendix Table C.5.

Table C.2: By believed accuracy: evaluators' posterior beliefs about workers in *Evaluator Study* when gender is known

	I accounted for gender differences in confidence:		
	Just right	Too much	Too little
	(1)	(2)	(3)
Δ	11.16***	12.74***	7.40*
	(2.29)	(4.26)	(4.12)
Δ*Attention	2.81	-8.93	-1.88
	(3.17)	(6.27)	(5.53)
Δ*Calculation	-5.61*	-4.70	-6.21
	(3.27)	(6.75)	(5.37)
N	761	169	280
Condition FE	yes	yes	yes
Truth(Δ)	1.74	1.74	1.74

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are robust and shown in parentheses. The data are from the *Baseline*, *Attention*, and *Calculation* treatments for the group of evaluators noted in the column, specifically evaluators who, in the follow-up survey, indicate that they: believe they accurately accounted in this study for any gender differences in confidence in Column 1, believe they accounted "too much" in this study for gender differences in confidence in Column 2, and believe they accounted "too much" in this study for gender differences in confidence in Column 3. The regression specifications are the same as in Appendix Table C.5.

Table C.3: By beliefs about employers: evaluators' posterior beliefs about workers in *Evaluator Study* when gender is known

	Employers account for gender differences in confidence:		
	Just right	Too much	Too little
	(1)	(2)	(3)
Δ	12.21***	5.38	12.23***
	(3.29)	(4.11)	(2.41)
Δ*Attention	-3.01	9.45*	-2.36
	(5.30)	(5.52)	(3.28)
Δ*Calculation	-0.14	-11.39**	-5.44
	(5.40)	(5.53)	(3.35)
N	247	283	680
Condition FE	yes	yes	yes
Truth(Δ)	1.74	1.74	1.74

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are robust and shown in parentheses. The data are from the *Baseline*, *Attention*, and *Calculation* treatments for the group of evaluators noted in the column, specifically evaluators who, in the follow-up survey, indicate that they believe that employers' hiring, pay and promotion decisions: "accurately account for" the gender gap in confidence in Column 1, "need to account more" for the gender gap in confidence in Column 2, and "account too much" for the gender gap in confidence in Column 3. The regression specifications are the same as in Appendix Table C.5.

Table C.4: By cognitive ability measures: evaluators' posterior beliefs about workers in *Evaluator Study* when gender is known

X:	Demeaned CRT score	Indicator for Base Rate Pure Neglect	Demeaned error in base rate questions	Demeaned error in Bayesian updating question
	(1)	(2)	(3)	(4)
Panel A: *Baseline* treatment				
Δ	8.64***	7.77***	8.76***	8.91***
	(1.79)	(2.15)	(1.78)	(1.78)
Δ*X	0.96	3.15	0.35**	-0.09
	(1.48)	(3.83)	(0.17)	(0.09)
N	402	402	402	402
Panel B: *Attention* treatment				
Δ	9.02***	6.55***	9.14***	9.11***
	(1.73)	(2.02)	(1.72)	(1.73)
Δ*X	0.90	7.91**	0.36**	-0.16*
	(1.47)	(3.86)	(0.14)	(0.09)
N	403	403	403	403
Panel C: *Calculation* treatment				
Δ	3.13*	2.14	3.18*	3.20*
	(1.79)	(2.08)	(1.81)	(1.80)
Δ*X	1.60	3.83	-0.04	-0.10
	(1.51)	(4.18)	(0.14)	(0.08)
N	405	405	405	405
Suppressed X	yes	yes	yes	yes
Truth(Δ)	1.74	1.74	1.74	1.74

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are robust and shown in parentheses. The data are from the *Baseline*, *Attention*, and *Calculation* treatments in Panels A, B, and C, respectively. Each column presents an OLS of evaluators' posterior beliefs on (i) an indicator for being asked about female workers (Δ), (ii) a (suppressed) measure of X, and (iii) an interaction of the indicator in (i) and the measure of X. X is noted in each column and is: an evaluator's demeaned CRT score (out of three questions) in Column 1, an indicator for whether the evaluator exhibited pure base rate neglect (where pure base rate neglect is consistent with ignoring the prior likelihood entirely) in Column 2, the demeaned distance between the evaluator's answer and the Bayesian posterior in the base rate neglect bonus question in Column 3, and the demeaned distance between the evaluator's answer and the Bayesian posterior in the Bayesian updating bonus question in Column 4. At the bottom of the table, we provide corresponding true values for the difference in evaluators' beliefs about female and male workers if evaluators are fully accurate when they are asked about female and male workers (see the estimates Truth(Δ)).

Table C.5: By demographics: evaluators' posterior beliefs about workers in *Evaluator Study* when gender is known

	Men	Women	Low Education	High Education	Low Income	High Income
	(1)	(2)	(3)	(4)	(5)	(6)
Δ	9.31***	11.56***	11.38***	9.73***	11.33***	9.94***
	(3.02)	(2.27)	(2.60)	(2.46)	(3.00)	(2.20)
Δ*Attention	-2.05	1.76	-0.93	1.54	-0.47	0.95
	(4.18)	(3.09)	(3.64)	(3.42)	(3.96)	(3.20)
Δ*Calculation	-4.31	-5.83*	-5.99	-5.66	-5.24	-5.93*
	(4.08)	(3.36)	(3.77)	(3.45)	(4.03)	(3.31)
N	507	669	573	637	531	679
Condition FE	yes	yes	yes	yes	yes	yes
Truth(Δ)	1.74	1.74	1.74	1.74	1.74	1.74

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are robust and shown in parentheses. The data are from the *Baseline*, *Attention*, and *Calculation* treatments for the group of evaluators noted in the column, specifically evaluators who: are men in Column 1, are women in Column 2, have an educational attainment of an Associate's Degree or less in Column 3, have an educational attainment of Bachelor's Degree or more in Column 4, have a reported annual income of below $50,000 in Column 5 and have a report annual income equal to or exceeding $50,000 in Column 6. Each column presents an OLS of evaluators' posterior beliefs on (i) suppressed indicators (i.e., Condition FEs) for the *Baseline*, *Attention*, and *Calculation* treatments as well as (ii) an indicator for being asked about female workers (Δ) and an indicator for being asked about female workers interacted with the indicator for the X treatment (Δ*X). At the bottom of the table, we provide corresponding true values for the difference in evaluators' beliefs about female and male workers if evaluators are fully accurate when they are asked about female and male workers (see the estimates Truth(Δ)).

Table C.6: By more demographics: evaluators' posterior beliefs about workers in *Evaluator Study* when gender is known

	Younger	Older	Favor Democrats	Favor Republicans
	(1)	(2)	(3)	(4)
Δ	9.19***	12.04***	9.78***	11.81***
	(2.32)	(2.74)	(2.15)	(3.17)
Δ*Attention	0.53	0.29	1.09	-0.79
	(3.28)	(3.80)	(2.99)	(4.53)
Δ*Calculation	-2.71	-9.37**	-4.96	-6.97
	(3.26)	(4.01)	(3.02)	(4.63)
N	691	519	826	384
Condition FE	yes	yes	yes	yes
Truth(Δ)	1.74	1.74	1.74	1.74

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are robust and shown in parentheses. The data are from the *Baseline*, *Attention*, and *Calculation* treatments for the group of evaluators noted in the column, specifically evaluators who: are 18-35 year old in Column 1, are 36 years or older in Column 2, indicate that they feel more favorably about Democrats than Republicans in Column 3, and indicate that they feel (weakly) more favorably about Republicans than Democrats. The regression specifications are the same as in Appendix Table C.5.

D Additional Robustness Results

In this Appendix, we present results from several additional study versions. See Section D.1 for the *Evaluator (Professional Evaluators) Study* and corresponding *Worker (Undergraduate Study) Study*, Section D.2 for the *Baseline* treatment of the *Evaluator (Extended) Study*, Section D.3 for the *Strategic Incentives* treatment of the *Evaluator (Extended) Study* and corresponding *Strategic Incentives* treatment of the *Worker Study*, Section D.4 for the *Joint Evaluations* and *Joint Evaluations, Strategic Incentives* treatments of the *Evaluator (Extended) Study*, Section D.5 for the *Evaluator (Alternative Questions) Study*, Section D.6 for the *Evaluator (Additional Demographics) Study*, and Section D.7 for the *Evaluator (Known Performance) Study*.

D.1 The *Baseline* and *Baseline, Unknown Gender* Treatments of The *Evaluator (Professional Evaluators) Study* and The *Worker (Undergraduates) Study*

We designed the *Evaluator (Professional Evaluators) Study* and corresponding *Worker (Undergraduates) Study* to investigate whether our results still hold with evaluators who have experience in management and hiring, and who are evaluating workers that they might be more familiar with.

To begin, we recruited 354 undergraduate students from Ohio State University to complete our *Worker (Undergraduates) Study*. After excluding 4 participants who neither identify as men nor women because we are under-powered to consider this group, this resulted in 350 workers. These workers take a similar 10-question math and science test and provide similar beliefs as the workers in our main *Worker Study*; see Appendix Table A.6 for a discussion of the minor differences between the *Worker (Undergraduates) Study* and *Worker Study*.

After recruiting these workers, we then recruited 800 professional evaluators for our *Evaluator (Professional Evaluators) Study*. Specifically, we use the internal screening questions on Prolific to recruit to the subset of Prolific users who answered "yes" to the following two questions: (i) Do you have any experience in making hiring decisions (i.e. have you been responsible for hiring job candidates)?, and (ii) Do you have any experience being in a management position? The instructions for the *Evaluator (Professional Evaluators) Study* were the same as the instructions for the *Baseline* treatment of the *Evaluator Study* with three notable expectations. First, we informed our professional evaluators that workers were undergraduate students from "a large Midwestern university who expected to graduate in Spring 2023." That is, our available pool of workers from the *Worker (Undergraduates) Study* is the group of workers who indicated that they expected to graduate in Spring 2023, which would be a natural pool of workers for our professional evaluators to consider. Second, the self-evaluation information that we provide to evaluators reflects the beliefs of these undergraduate students from the *Worker (Undergraduates) Study*. Third, rather than randomizing evaluators into one of 6 conditions, we randomize professional evaluators into either

the *Baseline* treatment or the *Baseline, Unknown Gender* treatment because of the limited sample size of professional evaluators given the associated screening criteria.

Appendix Table D.1 confirms that—for our main self-evaluation question—the confidence gap persists both for the overall study population and for the available pool of workers.[31] Most importantly, despite an insignificant performance gap of 1.91 percentage points among the available pool of workers, Column 3 of Appendix Table D.1 shows that there is a substantial and statistically significant confidence gap of 26.3 percentage points for our main self-evaluation: 58.6% of female workers believe they have a poor performance while only 32.3% of male workers believe they have a poor performance.

Table D.1: Self-Evaluations in the *Baseline* treatment of the *Worker (Undergraduates) Study*

	All Workers		Available Pool of Workers	
	(1)	(2)	(3)	(4)
Female	0.176***	0.121**	0.263**	0.222*
	(0.053)	(0.053)	(0.115)	(0.119)
Constant	0.394***		0.323***	
	(0.039)		(0.085)	
N	350	350	72	72
Perf FE	No	Yes	No	Yes

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are robust and shown in parentheses. Results are from OLS regressions of the responses provided to the main self-evaluation question, coded as 1 if the workers guess they have a "poor performance" and 0 otherwise. *Female* is an indicator for the worker identifying as a woman. Perf FEs are dummies for each possible performance out of the 10 questions on the test. In Columns 1–2, data are from the 350 participants who identified as a man or a woman in the *Baseline* Treatment of the *Worker (Undergraduates) Study*. In Columns 3–4, data are further restricted to the available pool of workers that evaluators are asked about—i.e., male and female workers who expect to graduate in 2023.

Appendix Table D.2 presents the results for the professional evaluators in the *Baseline* treatment. According to their priors, professional evaluators appear slightly more accurate than our main evaluators in terms of their expected performance gap. As shown in Column 1, professional evaluators expect an insignificant performance gap of 1.86 percentage points (see Panel A), which is very similar to and statistically indisguishabe from the true performance gap of 1.91 percentage points (see Panel B). Given these priors and that professional evaluators indeed expect a confidence gap according to their overconfidence beliefs (see Column 2) and underconfidence beliefs (Column 3), their implied Bayesian posteriors also indicate that—if they are Bayesian—being provided with the information on the workers' self-evaluation should *not* cause them to have more pessimistic be-

[31] Similar results follow from the other self-evaluation questions as well. Specifically, results this study replicate the confidence gap: out of the 13 self-evaluation questions they are asked, when controlling for performance fixed effects and considering all 350 workers, we find that women provide worse self-evaluations in response to all 13 questions and significantly so in response to 10 out of the 13 questions.

liefs about women (see Column 4). Nonetheless, just as with our main participants, the confidence gap conveyed via this information causes professional evaluators to form much more pessimistic beliefs about women. According to their posteriors, professional evaluators inaccurately expect a substantial and statistically significant performance gap of 14.65 (Column 5).

Appendix Table D.3 presents the results on evaluators' beliefs in the *Baseline, Unknown Gender*. We note that these results are very similar to those in the *Baseline* treatment and indeed the only significant differences that emerge are as follows: the overconfidence gap is significantly smaller now that gender is unknown and the underconfidence gap is marginally significantly smaller when gender is now unknown.

Thus, while our professional evaluators are more likely to expect the confidence gap when gender is known compared to when it is unknown, we find no evidence for professional evaluators being better able to account for the *gender* gap in confidence (as evident via their posterior beliefs in Column 5 of Appendix Tables D.2 and D.3).

Table D.2: Evaluators' Beliefs in the *Baseline* Treatment of the *Evaluator (Professional Evaluators) Study*

	Prior	Over-confidence	Under-confidence	Implied Bayesian Posterior	Posterior
	(1)	(2)	(3)	(4)	(5)
Panel A: Evaluators' Beliefs					
B(F)	37.87	38.78	52.64	38.60	50.37
B(M)	36.25	49.61	37.57	36.73	35.71
Δ	1.62	-10.83***	15.07***	1.87	14.65***
	(1.89)	(2.16)	(2.00)	(1.83)	(1.48)
Panel B: Evaluators' Beliefs - Truth					
B(F)	8.60	-1.79	-5.53	9.33	21.10
B(M)	8.90	5.88	14.35	9.38	8.36
Δ	-0.30	-7.67***	-19.88***	-0.05	12.73***
	(1.89)	(2.16)	(2.00)	(1.83)	(1.48)
N	409	409	409	406	409
Truth(F)	29.27	40.57	58.17	29.27	29.27
Truth(M)	27.35	43.73	23.22	47.79	27.35
Truth(Δ)	1.91	-3.16	34.95	1.91	1.91

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are robust and shown in parentheses. Results follow the structure of Table 2. Data are from the 409 participants in the *Baseline* treatment of the *Evaluator (Professional Evaluators) Study*. Sample size differs slightly in column (4) as some evaluators' beliefs imply a Bayesian posterior that is undefined.

Table D.3: Evaluators' Beliefs in the *Baseline, Unknown Gender* Treatment of the *Evaluator (Professional Evaluators) Study*

	Prior	Over-confidence	Under-confidence	Implied Bayesian Posterior	Posterior
	(1)	(2)	(3)	(4)	(5)
Panel A: Evaluators' Beliefs					
B(F)	39.25	42.87	49.78	40.22	50.46
B(M)	38.03	43.90	39.56	36.49	36.61
Δ	1.22	-1.02	10.22***	3.73*	13.84***
	(1.97)	(2.22)	(2.05)	(1.91)	(1.49)
Panel B: Evaluators' Beliefs - Truth					
B(F)	9.98	2.30	-8.39	10.95	21.19
B(M)	10.68	0.17	16.34	9.14	9.26
Δ	-0.70	2.14	-24.73***	1.81	11.92***
	(1.97)	(2.22)	(2.05)	(1.91)	(1.49)
N	391	391	391	391	391
Truth(F)	29.27	40.57	58.17	29.27	29.27
Truth(M)	27.35	43.73	23.22	47.79	27.35
Truth(Δ)	1.91	-3.16	34.95	1.91	1.91

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are robust and shown in parentheses. Results follow the structure of Table 2. Data are from the 391 participants in the *Unknown Gender* treatment of the *Evaluator (Professional Evaluators) Study*.

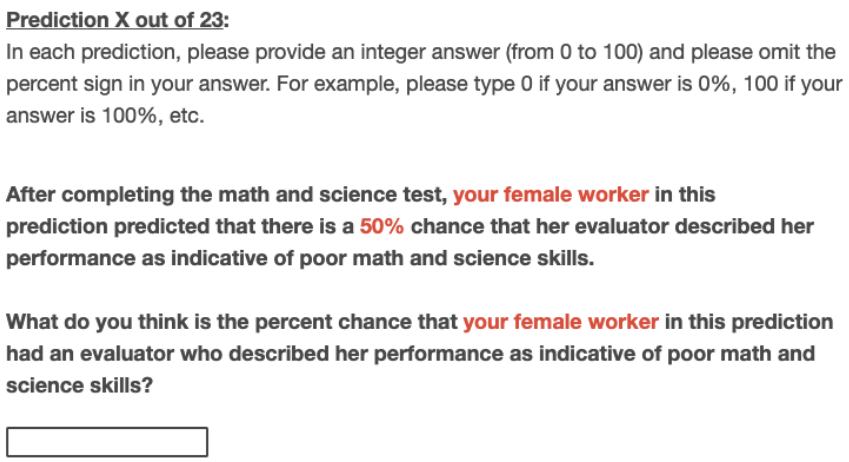
D.2 The *Evaluator (Extended) Study*

We ran the *Evaluator (Extended) Study* to test whether evaluators are better able to account for the confidence gap as conveyed via self-evaluations if they have more experience with how workers answer other self-evaluation questions.

We recruited 406 additional evaluators from Prolific. Evaluators first are asked to provide prior beliefs in the same manner as in the *Baseline* treatment of the *Evaluator Study*. Then, to gain experience with the self-evaluations of workers, evaluators are asked to make 20 predictions about specific workers after learning that specific workers' self-evaluation. Specifically, on each decision screen, evaluators are informed of a specific worker's reported percent chance of having a poor performance and then are asked to make a prediction about the percent chance of that specific worker having a poor performance (see Appendix Figure D.1 below).[32]

[32]While we provide evaluators with information on how these specific workers answer the continuous Self-Evaluation Question 8C, the aggregate information we provide about the workers' self-evaluations when eliciting our main posterior belief relates to how workers answered the binary Self-Evaluation Question 8B, consistent with our main *Evaluator Study*.

Figure D.1: Screenshot of Posterior Belief in *Baseline* Treatment of the *Evaluator (Extended) Study*



After evaluators provide these 20 worker-specific beliefs, we ask them for their posterior belief, overconfidence belief, and underconfidence belief in the same manner as in the *Baseline* treatment of the *Evaluator Study*. Thus, the only difference between the "experienced" evaluators in the *Evaluator (Extended) Study* and the evaluators in the main *Evaluator Study* is that the experienced evaluators have seen 20 additional worker-specific self-evaluations and have reported 20 corresponding worker-specific beliefs. Therefore, if gaining experience with the worker-specific self-evaluations helps evaluators to adjust for worker confidence, then we would expect to see the gender difference in posteriors reduced in the *Evaluator (Extended) Study*.

Appendix Table D.4 presents the results for these experienced evaluators. Experience does not help evaluators to better account for the gender gap in confidence. Among these experienced evaluators, their overconfidence and underconfidence beliefs indicate that they expect significant gender gaps in confidence, and their Bayesian posterior beliefs imply that they should—if they are Bayesian—expect little-to-no performance gap. However, according to their posteriors, even experienced evaluators expect a large performance gap (∼15 percentage point).

Appendix Figure D.2 and Appendix Table D.5 show how evaluators' beliefs respond to individual worker's self-evaluations, as discussed in Section 6.3. Appendix Figure D.2 shows that there is some evidence that evaluators account for the confidence gap among the most pessimistic self-evaluations. For instance, when a worker reports an 80% chance of having a poor performance in their self-evaluation, the average evaluator believes there is a 74% chance of that worker having a poor evaluation if the worker is a man but only a 70% of that worker having a poor evaluation if that worker is a woman. Nonetheless, Appendix Table D.5 shows that—even when asked about specific workers—evaluators expect a statistically significant performance gap (∼4.65 percentage points), according to their posterior beliefs.

Table D.4: Evaluators' Beliefs in the *Baseline* Treatment of the *Evaluator (Extended) Study*

	Prior	Over-confidence	Under-confidence	Implied Bayesian Posterior	Posterior
	(1)	(2)	(3)	(4)	(5)
Panel A: Evaluators' Beliefs					
B(F)	40.21	38.21	52.52	42.78	65.72
B(M)	38.35	45.91	43.46	39.70	50.97
Δ	1.86	-7.69***	9.05***	3.08*	14.75***
	(1.65)	(2.27)	(2.14)	(1.68)	(1.49)
Panel B: Evaluators' Beliefs - Truth					
B(F)	-9.32	22.86	-22.28	-6.75	16.19
B(M)	-9.44	6.85	-8.68	-8.09	3.18
Δ	0.12	16.02***	-13.61***	1.34	13.01***
	(1.65)	(2.27)	(2.14)	(1.68)	(1.49)
N	406	406	406	404	406
Truth(F)	49.53	15.35	74.80	49.53	49.53
Truth(M)	47.79	39.06	52.14	47.79	47.79
Truth(Δ)	1.74	-23.70	22.65	1.74	1.74

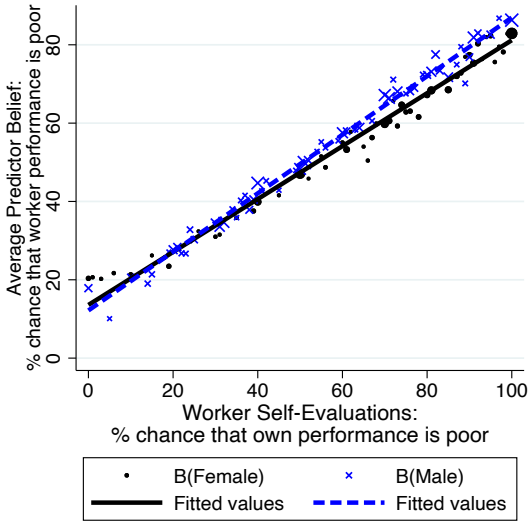
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are robust and shown in parentheses. Results follow the structure of Table 2. Data are from the 406 participants in the *Baseline* treatment of the *Evaluator (Extended) Study*. Sample size differs slightly in column (4) as some evaluators' beliefs imply a Bayesian posterior that is undefined.

Table D.5: Evaluators' Beliefs about Specific Workers in the *Baseline* treatment of the *Evaluator Study*

	(1)	(2)
Δ	4.65***	4.68***
	(1.11)	(1.11)
Constant	55.08***	
	(0.72)	
N	8120	8120
Performance FE	no	yes

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are clustered at the evaluator level. Results are from OLS regressions of the believed chance that a specific worker has a poor performance after learning that worker's self-evaluation (i.e., the percent chance that they believed they had a poor evaluation) on an indicator for being asked about female workers (Δ). Data are from the 20 observations for each of the 406 participants in the *Baseline* treatment of the *Evaluator (Extended) Study*.

Figure D.2: Evaluators' Beliefs About Specific Workers as a Function of Worker's Self-Evaluation



Graph shows a scatterplot of the average believed chance that a worker had a poor performance against that worker's believed percent chance that they had a poor performance. Data are from the *Evaluator (Extended) Study*.

D.3 The *Worker(Strategic Incentives) Study* and the *Evaluator (Extended, Strategic Incentives) Study*

We designed the *Strategic Incentives* treatment of the *Worker Study* and the *Strategic Incentives* treatment of the *Evaluator (Extended) Study* to see how evaluators update about workers who face strategic incentives.

Turning first to the workers, we recruit 387 new participants through Prolific. These workers face incentives that are akin to those in the *Self-Promotion* treatment of Exley and Kessler (2022). The workers are told that—if Part 2 is randomly selected as the part-that-counts—their "employer," who is another Prolific participant who completes the *Employer Study* (see footnote 33 for details on that study), will decide whether or not to hire them after only learning their answer in a randomly-selected self-evaluation. If they are not hired, then they will earn a bonus payment of \$0.50 and their employer will earn a bonus payment of \$0.50. If they are hired, then they will earn a bonus payment of \$1 and their employer will earn a bonus payment equal to \$0.10 times the number of questions they answered correctly on the math and science test.[33]

[33]We ran the *Employer Study* only to incentivize these decisions, so we do not present detailed results. In short summary, we recruited 100 Prolific participants to act as employers, and used a strategy method elicitation to ask whether they would hire their worker for each of the possible self-evaluations that the worker could have given in the 8 binary self-evaluation questions (Questions 1B, 2B, ..., 8B in Appendix Table A.1) and the possible absolute performance guesses that the worker could have given (Question 0 in Appendix Table A.1). Employers do not know workers' gender. We find that, for all binary self-evaluations, employers

Then, we recruited 394 additional participants as evaluators who are asked to make predictions about these workers who faced strategic incentives. Evaluators are informed of the incentives workers faced before they make their predictions.

Appendix Table D.6 presents results on these workers, as discussed in Section 6.4[34] In addition, we also note that the persistence of the confidence gap when workers face strategic incentives is *not* reflective of workers being unresponsive to strategic incentives. Rather, while strategic incentives cause both male and female workers to report significantly more favorable self-evaluations in response to the 13 out of the 17 self-evaluation questions, the gender difference in self-evaluations is statistically significant in 16 out of the 17 self-evaluations questions. This is because the impact of the strategic incentives is similar among men and women in response to all 17 self-evaluation questions—replicating another finding from Exley and Kessler (2022).

Appendix Table D.7 presents results on these evaluators, as discussed in Section 6.4. In addition, we note that these evaluators in the *Strategic Incentives* treatment are very similar to evaluators to evaluators in the *Baseline* treatment of the *Evaluator (Extended) Study*. Specifically, comparing results in these two treatments reveals no significant differences in the specifications shown in Columns 1–4 of Appendix Table D.7. Differences only emerge in Column 5. Evaluators in the Strategic Incentives treatment expect that male workers are slightly more likely to have poor performance and expect that gender difference in the likelihood of having a poor performance is slightly smaller.

are significantly more likely to hire workers if they provided a positive self-evaluation compared to a negative self-evaluation. Furthermore, a worker's chance of being hired is significantly increasing in their answer to the absolute performance self-evaluation. Thus, workers who provide more optimistic self-evaluations are more likely to be hired and therefore earn higher payments.

[34]Similar results follow from the other self-evaluation questions as well. Specifically, results this study replicate the confidence gap: out of the 17 self-evaluation questions they are asked, when controlling for performance fixed effects and considering all 387 workers, we find that women provide worse self-evaluations in response to all 17 questions and significantly so in response to 10 out of the 16 questions.

Table D.6: Self-Evaluations in the *Strategic Incentives* treatment of the *Worker Study*

	All Workers		Available Pool of Workers	
	(1)	(2)	(3)	(4)
Female	0.194***	0.168***	0.173***	0.160***
	(0.049)	(0.048)	(0.059)	(0.059)
Constant	0.510***		0.567***	
	(0.036)		(0.044)	
N	387	387	250	250
Perf FE	No	Yes	No	Yes

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are robust. Results are from OLS regressions of the responses provided to the main self-evaluation question, coded as 1 if the workers guess they have a "poor performance" and 0 otherwise. *Female* is an indicator for the worker identifying as a woman. Perf FEs are dummies for each possible performance out of the 10 questions on the test. In Columns 1–2, data are from the 387 participants who identified as a man or a woman in the *Strategic Incentives* Treatment of the *Worker Study*. In Columns 3–4, data are further restricted to the available pool of workers that evaluators are asked about—i.e., male and female workers with performances in the "middle" or 25th-75th percentile.

Table D.7: Evaluators' Beliefs' about Workers in the *Strategic Incentives* treatment of the *Evaluator (Extended) Study*

	Prior	Over-confidence	Under-confidence	Implied Bayesian Posterior	Posterior
	(1)	(2)	(3)	(4)	(5)
Panel A: Evaluators' Beliefs					
B(F)	40.55	38.45	55.80	41.41	62.92
B(M)	39.45	47.22	43.14	41.15	53.77
Δ	1.09	-8.77***	12.66***	0.26	9.16***
	(1.71)	(2.22)	(2.03)	(1.65)	(1.31)
Panel B: Evaluators' Beliefs - Truth					
B(F)	-10.42	12.86	-17.75	-9.56	11.95
B(M)	-10.08	10.07	-7.51	-8.38	4.24
Δ	-0.35	2.79	-10.24***	-1.18	7.72***
	(1.71)	(2.22)	(2.03)	(1.65)	(1.31)
N	394	394	394	394	394
Truth(F)	50.97	25.59	73.55	50.97	50.97
Truth(M)	49.53	37.15	50.65	49.53	49.53
Truth(Δ)	1.44	-11.56	22.89	1.44	1.44

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are robust and shown in parentheses. Results follow the structure of Table 2. Data are from the 394 participants in the *Strategic Incentives* treatment of the *Evaluator (Extended) Study*.

D.4 The *Joint Evaluations* and *Joint Evaluations, Strategic Incentives* Treatments of the *Evaluator (Extended) Study*

We ran the *Joint Evaluations* treatment of the *Evaluator (Extended) Study* to investigate whether evaluators are better able to account for the confidence gap when they simultaneously provide beliefs about male and female workers. The *Joint Evaluations* treatment follows the *Baseline* treatment of the *Evaluator (Extended) Study* except that each evaluator provides beliefs about male and female workers simultaneously, rather than providing beliefs about only one randomly-selected group of workers.

Specifically, we recruited 205 additional participants through Prolific to act as evaluators. Evaluators first report—on the same decision screen—their prior belief that a randomly selected male worker has a poor performance and that a randomly selected female worker has a poor performance. Then, evaluators are asked to make 20 predictions about individual workers. Rather than seeing 20 individual workers on their own decision screen, evaluators see one male worker and one female worker on the same screen and report their prediction for these workers simultaneously, and repeat this process ten times for a total of 20 workers. Then, evaluators see the percent of male workers and female workers who believed they have a poor performance, provide their posterior beliefs, and then provide their over- and underconfidence beliefs, again about male and female workers on the same decision screen.

Appendix Table D.8 presents the results for these evaluators, as discussed in Section 6.5. Joint evaluation does not eliminate the expected performance gap: these evaluators have expect a large and statistically significant performance gap ($\sim$15 percentage point), according to their posteriors.

In addition, by leveraging the fact that these evaluators are asked about both men and women, Appendix Table D.9 allows us to further show that our results even persist among evaluators with incentivized overconfidence beliefs that indicate that they believe men are more overconfident than women (conditional on poor performance) and among evaluators with incentivized underconfidence beliefs that indicate that they believe women are more underconfident than men (conditional on good performance).

Finally, we also recruited an additional 195 Prolific participants and ran a *Joint Evaluations, Strategic Incentives* treatment of the *Evaluator (Extended) Study*. This is the same as the *Joint Evaluations* treatment described above, except evaluators were matched with the workers from the *Worker (Strategic Incentives) Study* described in Appendix Section D.3. Appendix Table D.10 presents results from these evaluators; results are similar to above.

Table D.8: Evaluators' Beliefs' about Workers in the *Joint Evaluations* treatment of the *Evaluator (Extended) Study*

	Prior	Over-confidence	Under-confidence	Implied Bayesian Posterior	Posterior
	(1)	(2)	(3)	(4)	(5)
Panel A: Evaluators' Beliefs					
B(F)	41.79	31.51	54.40	44.90	68.18
B(M)	38.80	49.96	34.40	41.79	53.45
Δ	2.99**	-18.45***	20.00***	3.11**	14.73***
	(1.51)	(2.17)	(2.14)	(1.54)	(1.27)
Panel B: Evaluators' Beliefs - Truth					
B(F)	-7.74	16.16	-20.40	-4.63	18.65
B(M)	-8.99	10.90	-17.74	-6.00	5.66
Δ	1.25	5.26**	-2.66	1.37	12.99***
	(1.51)	(2.17)	(2.14)	(1.54)	(1.27)
N	410	410	410	408	410
Truth(F)	49.53	15.35	74.80	49.53	49.53
Truth(M)	47.79	39.06	52.14	47.79	47.79
Truth(Δ)	1.74	-23.70	22.65	1.74	1.74

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are robust and shown in parentheses. Results follow the structure of Table 2. Data are from the 410 participants in the *Joint Evaluations* treatment of the *Evaluator (Extended) Study*. Sample size differs slightly in column (4) as some evaluators' beliefs imply a Bayesian posterior that is undefined.

Table D.9: According to Evaluators' Overconfidence and Underconfidence Beliefs, Evaluators Posterior Beliefs' about Workers in the *Joint Evaluations* treatment of the *Evaluator (Extended) Study*

Belief:	Men are more overconfident	Men are NOT more overconfident	Women are more underconfident	Women are NOT more underconfident
	(1)	(2)	(3)	(4)
B(F)	67.35	70.76	67.75	71.15
B(M)	54.17	51.22	53.88	50.50
Δ	13.18***	19.54***	13.87***	20.65***
	(1.47)	(2.52)	(1.38)	(3.08)
Constant	54.17***	51.22***	53.88***	50.50***
	(0.85)	(1.48)	(0.80)	(1.97)
N	310	100	358	52

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are robust and shown in parentheses. Results follow the specification of Column 5 in Table 2. Columns 1 -4 restrict to the set of participants, who given their overconfidence and underconfidence beliefs about men and women, believe that (i) men are more overconfident conditional on poor performance in Column 1, (ii) men are not more overconfident conditional on poor performance in Column 2, (iii) women are more underconfident conditional on good performance in Column 3, and (iv) women are not more underconfident conditional on good performance in Column 4. Data are from the 410 participants in the *Joint Evaluations* treatment of the *Evaluator (Extended) Study*.

Table D.10: Evaluators' Beliefs' about Workers in the *Joint Evaluations, Strategic Incentives* treatment of the *Evaluator (Extended) Study*

	Prior	Over-confidence	Under-confidence	Implied Bayesian Posterior	Posterior
	(1)	(2)	(3)	(4)	(5)
Panel A: Evaluators' Beliefs					
B(F)	41.05	33.84	55.91	42.85	62.75
B(M)	38.46	51.50	35.03	41.21	51.81
Δ	2.58	-17.66***	20.89***	1.65	10.94***
	(1.59)	(2.15)	(2.06)	(1.52)	(1.19)
Panel B: Evaluators' Beliefs - Truth					
B(F)	-9.92	8.25	-17.64	-8.12	11.78
B(M)	-11.07	14.35	-15.62	-8.32	2.28
Δ	1.14	-6.10***	-2.01	0.21	9.50***
	(1.59)	(2.15)	(2.06)	(1.52)	(1.19)
N	390	390	390	385	390
Truth(F)	50.97	25.59	73.55	50.97	50.97
Truth(M)	49.53	37.15	50.65	49.53	49.53
Truth(Δ)	1.44	-11.56	22.89	1.44	1.44

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are robust and shown in parentheses. Results follow the structure of Table 2. Data are from the 390 participants in the *Joint Evaluations, Strategic Incentives* treatment of the *Evaluator (Extended) Study*. Sample size differs slightly in column (4) as some evaluators' beliefs imply a Bayesian posterior that is undefined.

D.5 The *Evaluator (Alternative Questions) Study*

In our main evaluator results, we focus on evaluators who were asked to guess the percent chance that a worker has a poor performance. We focused on this question because subjective assessments are common in self-evaluations and performance reviews and these subjective assessments also allow us to test the robustness of our results to an environment in which evaluators know objective performance metrics of the workers (see Section 6.9). However, to assess the robustness of our results to beliefs about objective categories, we ran an additional study called the *Evaluator (Alternative Questions) Study*.

For this study, we recruited 400 new evaluators through Prolific. The *Evaluator (Alternative Questions) Study* was the same as the *Baseline* treatment of the *Evaluator* Study except that evaluators were asked to provide all of the beliefs listed in Appendix Table A.3 in addition to the beliefs listed in Appendix Table A.2. Four of these questions involve objective outcomes: the likelihood that a worker got more than 3 questions right out of 10, the likelihood that a worker got more than 5 questions right out of 10, and the likelihood that a worker got more than 7 questions right out of 10. Additionally, we asked valuators to predict the likelihood that a worker scored in the top half relative to other workers, and the likelihood of a "poor" performance when using alternative subjective poor performance definition (see "poor-2" belief questions shown in Appendix Table A.3).

Recall that, as discussed in Section 3 and presented in Appendix Table B.1, female workers report more pessimistic self-evaluations in all of these alternative questions, and significantly so in all but one question. Our main question in the *Evaluator (Alternative Questions) Study* is whether the expected performance gap persists in evaluators' beliefs about these alternative performance outcomes.

Appendix Table D.11 presents these results, as discussed in Section 6.7. We find that—directionally, and almost always at a statistically significant level—our results hold across all of these performance outcomes: evaluators' priors indicate little to no gender differences, evaluators expect that male workers are more likely to be overconfident and female workers are more likely to be underconfident, their implied Bayesian beliefs imply no expected performance gap, but their posteriors indicate large and significant expected performance gaps.

Table D.11: Evaluators' Beliefs in the *Evaluator (Alternative Questions) Study*

	Belief Before Self-Eval Info	Over-confidence	Under-confidence	Implied Bayesian Belief	Belief After Self-Eval Info
	(1)	(2)	(3)	(4)	(5)
Panel A: Beliefs (main self-evaluation) about poor performance					
B(F)	36.86	38.20	51.86	40.23	56.18
B(M)	40.98	49.93	46.60	41.70	49.67
Δ	-4.11**	-11.73***	5.25**	-1.47	6.51***
	(1.68)	(2.21)	(2.15)	(1.70)	(1.74)
Panel B: Beliefs (poor-2) about poor performance using alternative subjective definition					
B(F)	36.67	37.76	53.55	38.98	57.79
B(M)	38.55	51.07	48.24	39.71	51.61
Δ	-1.89	-13.31***	5.31**	-0.74	6.18***
	(1.76)	(2.26)	(2.14)	(1.76)	(1.82)
Panel C: Beliefs (3+) about 3+ questions right					
B(F)	76.85	40.32	49.97	75.51	76.61
B(M)	78.15	47.23	47.28	76.58	81.54
Δ	-1.30	-6.92**	2.69	-1.07	-4.93***
	(1.70)	(2.93)	(2.86)	(1.93)	(1.42)
Panel D: Beliefs (5+) about 5+ questions right					
B(F)	65.02	40.23	48.10	61.37	42.80
B(M)	62.07	49.59	45.99	61.01	51.50
Δ	2.95	-9.36***	2.11	0.36	-8.70***
	(1.87)	(2.24)	(2.14)	(1.89)	(1.68)
Panel E: Beliefs (7+) about 7+ questions right					
B(F)	49.82	42.27	51.30	47.65	22.43
B(M)	46.62	50.01	47.75	47.50	22.83
Δ	3.20	-7.74***	3.54	0.15	-0.40
	(2.21)	(2.74)	(2.50)	(2.56)	(1.97)
Panel F: Beliefs (top-half) about performed in the top-half					
B(F)	49.49	40.96	51.54	49.07	38.36
B(M)	48.98	51.00	46.54	49.82	47.99
Δ	0.52	-10.04***	5.00**	-0.75	-9.63***
	(1.81)	(2.30)	(2.18)	(1.80)	(1.49)
N	400	400	400	394	400

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are robust. Results are from OLS regressions of the same specifications as noted in Table 2. Panel A restricts to beliefs relating to the main self-evaluation question. Panels B–F restrict to beliefs relating to the additional self-evaluation questions as defined in Appendix Table A.3. Data are from the 400 participants in the *Evaluator (Alternative Questions) Study*. See Appendix Tables A.2 and A.3 for details on how these beliefs are elicited. Sample size differs slightly in column (4) as some evaluators' beliefs imply a Bayesian posterior that is undefined.

D.6 The *Evaluator (Additional Demographics) Study*

For the *Evaluator (Additional Demographics) Study*, we recruited 198 new evaluators and (truthfully) told these evaluators that their worker will be randomly drawn from a group of workers who work full time, are between 26 and 40 years old, live in the Southern region of the United States, have completed at least some college education, and are (wo)men.[35] Thus, gender is more subtly conveyed because it is only one of several demographic characteristics provided to evaluators. When comparing these groups of male and female workers, the female workers are—if anything—less likely to have a poor performance than male workers.[36] Nevertheless, as with our prior result, these female workers report significantly more pessimistic self-evaluations 77% of female workers in this group believe they have a poor performance while only 38% of male workers do.

Appendix Table D.12 presents these results, as discussed in Section 6.8. We find a very similar pattern of results: evaluators have posterior beliefs that indicate a large and statistically significant (∼23 percentage point) expected performance gap.

Table D.12: Evaluators' Beliefs' in the *Evaluator (Additional Demographics) Study*

	Prior	Over-confidence	Under-confidence	Implied Bayesian Posterior	Posterior
	(1)	(2)	(3)	(4)	(5)
Panel A: Evaluators' Beliefs					
B(F)	44.00	43.14	51.01	45.13	63.16
B(M)	41.43	48.15	39.67	41.10	40.52
Δ	2.57	-5.01	11.34***	4.03	22.65***
	(2.45)	(3.20)	(2.89)	(2.52)	(2.13)
Panel B: Evaluators' Beliefs - Truth					
B(F)	8.65	32.79	-18.96	9.78	27.81
B(M)	-1.69	-14.48	2.07	-2.02	-2.60
Δ	10.34***	47.27***	-21.03***	11.80***	30.42***
	(2.45)	(3.20)	(2.89)	(2.52)	(2.13)
N	198	198	198	198	198
Truth(F)	35.35	10.35	69.97	35.35	35.35
Truth(M)	43.12	62.63	37.60	43.12	43.12
Truth(Δ)	-07.77	-52.27	32.37	-07.77	-07.77

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are robust and shown in parentheses. Results follow the structure of Table 2. Data are from the 198 participants in the *Evaluator (Additional Demographics) Study*.

[35] These demographics were modal in the *Worker Study*, with modal age being the modal generation.
[36] Male versus workers have a 43% versus 35% of a poor performances ($p = 0.51$).

D.7 The *Evaluator (Known Performance) Study*

We ran our *Evaluator (Known Performance) Study* to investigate whether our results are robust to situations where more information is known about the quality of workers. Since our main self-evaluation question involves a subjective measure of performance, we can present a rather stringent test of whether our results are robust to a situation where worker quality is better known. Specifically, we can inform evaluators of a worker's *objective* performance and then investigate how the evaluators update about a *subjective* measure of the worker's performance.

In the *Evaluator (Known Performance) Study*, we recruited 198 new evaluators through Prolific. These evaluators are told that their worker will be randomly drawn from the group of male or female workers who got 5 questions right on the math and science test. Then, as in the main *Evaluator Study*, evaluators provide beliefs about whether their worker has a poor performance, which is equivalent to asking the evaluator to provide beliefs about whether a classifier—who is never informed of a worker's gender—believes a performance of 5 is poor.

Appendix Table D.13 presents these results, as discussed in Section 6.9. Even when evaluators are given precise information about a worker's quality, the self-evaluation information causes a large and significant (∼14 percentage points) expected performance gap in our subjective outcome.

Table D.13: Evaluators' Beliefs' in the *Evaluator (Known Performance) Study*

	Prior	Over-confidence	Under-confidence	Implied Bayesian Posterior	Posterior
	(1)	(2)	(3)	(4)	(5)
Panel A: Evaluators' Beliefs					
B(F)	41.10	44.50	53.68	41.30	58.29
B(M)	41.57	47.44	46.20	41.10	44.44
Δ	-0.46	-2.94	7.48***	0.20	13.85***
	(3.38)	(3.04)	(2.62)	(3.30)	(2.52)
Panel B: Evaluators' Beliefs - Truth					
B(F)	1.41	12.36	-14.18	1.61	18.60
B(M)	1.88	-11.38	5.02	1.41	4.75
Δ	-0.46	23.74***	-19.20***	0.20	13.85***
	(3.38)	(3.04)	(2.62)	(3.30)	(2.52)
N	198	198	198	198	198
Truth(F)	39.69	32.14	67.86	39.69	39.69
Truth(M)	39.69	58.82	41.18	39.69	39.69
Truth(Δ)	0.00	-26.68	26.68	0.00	0.00

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are robust and shown in parentheses. Results follow the structure of Table 2. Data are from the 198 participants in the *Evaluator (Known Performance) Study*.

E Bayesian Calculations

We calculate the Implied Bayesian Beliefs for two different types of outcomes: "poor" performances and "good" performances. We define "poor performance" and "good performance" separately for each specific performance outcome. Our poor performance outcomes are having a classifier who described the worker's performance as indicative of poor math and science skills (corresponding to Worker Question 7B and the main Evaluator questions), or having a classifier who described the worker's performance as poor (corresponding to Worker Question 8B and Evaluator Question poor-2 in the *Evaluator (Extended) Studies*). Our good performance outcomes all come from our *Evaluator (Extended) Studies*, and include getting 3 or more questions right (Worker Question 1B and Evaluator Question 3+), getting 5 or more questions right (Worker Question 2B and Evaluator Question 5+), getting 7 or more questions right (Worker question 3B and Evaluator Question 7+), and scoring in the top half when compared to other participants (Worker Question 4B and Evaluator Question Top Half).

In the following two subsections, we show how we calculate the Implied Bayesian Belief for these outcomes. For simplicity, we refer to all poor performance outcomes under the umbrella term "poor performance," and we refer to all good performance outcomes under the umbrella term "good performance."

E.1 Implied Bayesian Belief of Poor Performance

First, let us consider the main self-evaluation question and other "poor performance" outcomes. We say that the worker had a poor performance when they meet the classification of the poor performance metric. For example, in our main study, a worker had poor performance—which we denote here by *Poor*—if their classifier described their performance as indicative of poor math and science skills. In this case, a worker had a good performance—which we denote here by *Good*—if their classifier did not describe their performance as indicative of poor math and science skills. We say that a worker had a good self-evaluation (SE^{Good}) if the worker believed that they had a good performance, and a worker had a poor self-evaluation (SE^{Poor}) if the worker believed that they had a poor performance. For the main self-evaluation question, SE^{Good} corresponds to the worker believing that their classifier did not describe their performance as indicative of poor math and science skills and SE^{Poor} corresponds to the worker believing that their classifier described their performance as indicative of poor math and science skills. The definitions follow similarly for other poor performance outcomes.

We elicit the following beliefs from evaluators, where these beliefs refer to a randomly-selected

worker:

$$P(Poor) \equiv \% \text{ chance that the worker had a poor performance}$$
$$P(SE^{Poor}|Good) \equiv \% \text{ chance that the worker had a poor self-evaluation given that they had a good performance}$$
$$P(SE^{Good}|Poor) \equiv \% \text{ chance that the worker had a good self-evaluation given that they had a poor performance}$$

In the paper, we refer to $P(Poor)$ as the "prior belief," $P(SE^{Poor}|Good)$ as the "underconfidence belief," and $P(SE^{Good}|Poor)$ as the "overconfidence belief." The beliefs above imply the following "implied Bayesian posterior":

$\gamma_i \equiv \%$ chance that the worker had a poor performance, given that X% of workers had poor self-evaluations

To see this:

$$\begin{aligned}
\gamma_i &= P(Poor|X\%\ SE^{Poor}) \\
&= X\% * (P(Poor|SE^{Poor})) + (1 - X\%) * (P(Poor|SE^{Good})) \\
&= X\% * (1 - \underbrace{P(Good|SE^{Poor})}_{A}) + (1 - X\%) * \underbrace{P(Poor|SE^{Good})}_{B} \\
&= X * (1 - A) + (1 - X) * B
\end{aligned}$$

We can rewrite (A) into known terms as follows:

$$\begin{aligned}
(A) &= P(Good|SE^{Poor}) \\
&= \frac{P(Good \cap SE^{Poor})}{P(SE^{Poor})} \\
&= \frac{P(Good) * P(SE^{Poor}|Good)}{P(Good) * P(SE^{Poor}|Good) + (1 - P(Good)) * P(SE^{Poor}|Poor)} \\
&= \frac{(1 - P(Poor)) * P(SE^{Poor}|Good)}{(1 - P(Poor)) * P(SE^{Poor}|Good) + P(Poor) * (1 - P(SE^{Good}|Poor))} \\
&= \frac{(1 - \text{prior belief}) * \text{underconfidence belief}}{(1 - \text{prior belief}) * \text{underconfidence belief} + \text{prior belief} * (1 - \text{overconfidence belief})}
\end{aligned}$$

We can rewrite (B) into known terms as follows:

$$
\begin{aligned}
(B) &= P(Poor|SE^{Good}) \\
&= \frac{P(Poor \cap SE^{Good})}{P(SE^{Good})} \\
&= \frac{P(Poor) * P(SE^{Good}|Poor)}{P(Poor) * P(SE^{Good}|Poor) + (1 - P(Poor)) * P(SE^{Good}|Good)} \\
&= \frac{\text{prior belief} * \text{overconfidence belief}}{\text{prior belief} * \text{overconfidence belief} + (1 - \text{prior belief}) * (1 - \text{underconfidence belief})}
\end{aligned}
$$

E.2 Bayes of Good Performance

Now, let us consider the "good performance" outcomes. We say that the worker had a good performance when they meet the classification of the good performance metric. For example, a worker had a good performance—which we denote here by *Good*—if they got 3 or more questions right on the test. In this case, a worker had a poor performance—which we denote here by *Poor*—if they got fewer than 3 questions right. We say that the worker had a good self-evaluation (SE^{Good}) if the worker believed that they had a good performance, and a worker had a poor self-evaluation (SE^{Poor}) if the worker believed that they had a poor performance. For example, for self-evaluation Question 1B, SE^{Good} corresponds to the worker believing that they got 3 or more questions right on the test, and SE^{Poor} corresponds to the worker believing that they got fewer than 3 questions right on the test. The definitions follow similarly for the other good performance outcomes.

We elicit the following beliefs from evaluators, where these beliefs refer to a randomly-selected worker:

$$
\begin{aligned}
P(Good) &\equiv \% \text{ chance that the worker had a good performance} \\
P(SE^{Poor}|Good) &\equiv \% \text{ chance that the worker had a poor self-evaluation given that they had a good performance} \\
P(SE^{Good}|Poor) &\equiv \% \text{ chance that the worker had a good self-evaluation given that they had a poor performance}
\end{aligned}
$$

In the paper, for the good performance outcomes, we refer to $P(Good)$ as the "prior belief," "$P(SE^{Poor}|Good)$ as the "underconfidence belief," and $P(SE^{Good}|Poor)$ as the "overconfidence belief." The beliefs above imply the following "implied Bayesian posterior";

$\gamma_i \equiv$ % chance that a worker had a good performance, given that X% of workers had good self-evaluations

To see this:

$$\begin{aligned}
\gamma_i &= P(Good|X\%\ SE^{Good}) \\
&= X\% * (P(Good|SE^{Good})) + (1 - X\%) * (P(Good|SE^{Poor})) \\
&= X\% * (1 - \underbrace{P(Poor|SE^{Good})}_{A}) + (1 - X\%) * \underbrace{P(Good|SE^{Poor})}_{B} \\
&= X * (1 - A) + (1 - X) * B
\end{aligned}$$

We can rewrite (A) into known terms as follows:

$$\begin{aligned}
(A) &= P(Poor|SE^{Good}) \\
&= \frac{P(Poor \cap SE^{Good})}{P(SE^{Good})} \\
&= \frac{P(Poor) * P(SE^{Good}|Poor)}{P(Poor) * P(SE^{Good}|Poor) + (1 - P(Poor)) * P(SE^{Good}|Good)} \\
&= \frac{(1 - P(Good)) * P(SE^{Good}|Poor)}{(1 - P(Good)) * P(SE^{Good}|Poor) + P(Good) * (1 - P(SE^{Poor}|Good))} \\
&= \frac{(1 - \text{prior belief}) * \text{overconfidence belief}}{(1 - \text{prior belief}) * \text{overconfidence belief} + \text{prior belief} * (1 - \text{underconfidence belief})}
\end{aligned}$$

We can rewrite (B) into known terms as follows:

$$\begin{aligned}
(B) &= P(Good|SE^{Poor}) \\
&= \frac{P(Good \cap SE^{Poor})}{P(SE^{Poor})} \\
&= \frac{P(Good) * P(SE^{Poor}|Good)}{P(Good) * P(SE^{Poor}|Good) + (1 - P(Good)) * P(SE^{Poor}|Poor)} \\
&= \frac{P(Good) * P(SE^{Poor}|Good)}{P(Good) * P(SE^{Poor}|Good) + (1 - P(Good)) * (1 - P(SE^{Good}|Poor))} \\
&= \frac{\text{prior belief} * \text{underconfidence belief}}{\text{prior belief} * \text{underconfidence belief} + (1 - \text{prior belief}) * (1 - \text{overconfidence belief})}
\end{aligned}$$

E.3 Chance of Being Overconfident (Underconfident) Conditional on Bad (Good) Performance

Here, we derive the empirical probabilities of the likelihood that a randomly-selected worker is overconfident given poor performance or underconfident given good performance.

Following the definitions above, we define a good performance ($Good_i$) as worker i having been matched with a classifier who described their performance as good, and we define a poor performance ($Poor_i$) as worker i having been matched with a classifier who described their performance as poor.

Let's also define a good self-evaluation (SE_i^{Good}) as worker i indicating that they believe they were matched with a classifier who described their performance as good—hence believing that they had a good performance. Similarly, we define a poor self-evaluation (SE_i^{Poor}) as worker i indicating that they believe they were matched with a classifier who described their performance as poor—hence believing that they had a poor performance.

Given that classifiers were randomly assigned to workers, we say that worker i's chance of a poor performance—or their chance of having a classifier who denoted their performance as poor—is the chance that a randomly-selected classifier described worker i's performance as poor. This is analogous to the percent of classifiers who described i's score as a poor performance. We denote worker i's chance of a poor performance by $P(Poor)_i$.

To calculate the percent chance that a randomly-selected worker was overconfident given a poor performance, denoted $P(SE^{Good}|Poor)$, we note that:

$$P(SE^{Good}|Poor) = \frac{P(SE^{Good}) * P(Poor|SE^{Good})}{P(Poor)} \tag{1}$$

To determine the denominator of Equation 1, we note that $P(Poor)$, the probability that a randomly selected worker has a poor performance, is the chance of a worker having a poor performance, $P(Poor)_i$, averaged over all workers i. That is, if we index all workers from 1 to N:

$$P(Poor) = \frac{1}{N}\sum_{i}^{N} P(Poor)_i \tag{2}$$

Similarly, to determine the numerator of Equation 1, we note that:

$$P(SE^{Good}) * P(Poor|SE^{Good}) = \frac{1}{N}\sum_{i}^{N} P(SE_i^{Good}) * P(Poor|SE^{Good})_i \tag{3}$$

Then, we can plug in 2 and 3 to solve Equation 1 as follows:

$$P(SE^{Good}|Poor) = \frac{\frac{1}{N}\sum_{i}^{N} P(SE_i^{Good}) * P(Poor|SE^{Good})_i}{\frac{1}{N}\sum_{i}^{N} P(Poor)_i}$$

Since $P(SE_i^{Good})$ corresponds to individual i's binary guess of whether they had a good performance or not, this simply equals 0 or 1 for each worker i, and workers with a poor self-evaluation drop out of the numerator. Thus, this reduces to

$$P(SE^{Good}|Poor) = \frac{\sum_i^N P(Poor)_i * \mathbb{1}(SE_i^{Good} = 1)}{\sum_i^N P(Poor)_i} \tag{4}$$

Similarly, we solve $P(SE^{Poor}|Good)$ as follows

$$\begin{aligned} P(SE^{Poor}|Good) &= \frac{\sum_i^N P(Good)_i * \mathbb{1}(SE_i^{Poor} = 1)}{\sum_i^N P(Good)_i} \\ P(SE^{Poor}|Good) &= \frac{\sum_i^N (1 - P(Poor)_i) * \mathbb{1}(SE_i^{Poor} = 1)}{\sum_i^N (1 - P(Poor)_i)} \end{aligned} \tag{5}$$

Then, since we can calculate $P(Poor)_i$ for all worker i as the percent of evaluators who classify their performance as poor, and since we know whether each worker had a poor self-evaluation ($\mathbb{1}(SE_i^{Poor} = 1)$) or a good self-evaluation ($\mathbb{1}(SE_i^{Good} = 1)$), we can calculate Equations 4 and 5.

E.4 Bayesian Posterior Beliefs As A Function of Confidence

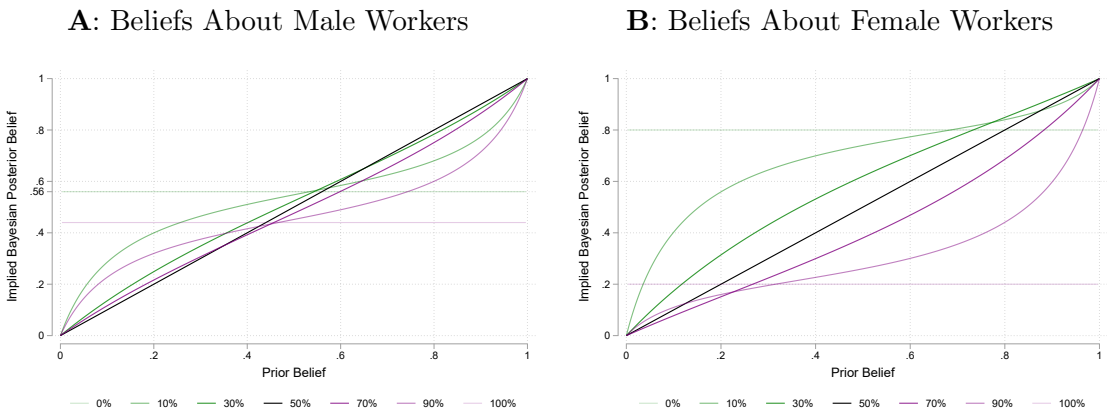
Appendix Figure E.1 shows how the levels of overconfidence and underconfidence beliefs affect the implied Bayesian posterior belief. These graphs plot the equation from Appendix Section E.1 as a function of the prior belief, overconfidence belief, and underconfidence belief. Panel A shows the implied Bayesian posterior belief for male workers, across the range of possible prior beliefs, for seven different example values of over- and underconfidence beliefs. Panel B shows the same but for female workers. For simplicity, we set the level of overconfidence belief equal to the level of underconfidence belief. The difference between the two panels lies in the signal that evaluators receive about workers. In particular, they are either given the signal that 56% of male workers believe that they have a poor performance, or they are given the signal that 80% of female workers believe that they have a poor performance. In a Bayesian framework, evaluators' over- and underconfidence beliefs affect how *informative* they believe this signal to be.

There are a few things evident from Appendix Figure E.1. First, if evaluators were to believe that workers are perfectly calibrated—that is, there is a 0% chance that workers are overconfident and a 0% chance that they are underconfident—the implied Bayesian posterior should be equal to the signal (56% for male workers and 80% for female workers) for all prior beliefs. This is the extreme in which evaluators believe that the signal is perfectly informative.[37] On the other extreme, over- and underconfidence beliefs of 50% correspond to a perfectly uninformative signal. In this case,

[37] On the other hand, when evaluators believe that there is a 100% chance that workers are over- or underconfident, the prior should be equal to one minus the signal.

the implied Bayesian posterior belief should be equal to the prior for all prior beliefs. As over- and underconfidence beliefs increase away from 0% toward 50%, the implied Bayesian posterior beliefs move toward the perfectly uninformative posterior. As as example shown in Appendix Figure E.1, when evaluators believe that there's a 30% chance that workers are over- and underconfident, the implied Bayesian posterior beliefs are already quite close to the perfectly uninformative benchmark.

Figure E.1: Implied Bayesian Posterior Beliefs as a Function of Prior Beliefs and Confidence



Graphs show the implied Bayesian posterior, across priors, for the overconfidence and underconfidence beliefs noted in the legend (assuming, for simplicity, that the level of the overconfidence and underconfidence belief is the same). Bayesian updating is done separately for male workers and female workers based on the actual signal given to evaluators. When updating about male workers, evaluators are told that 56% of male workers believed that they had a poor performance. When updating about female workers, evaluators are told that 80% of female workers believed that they had a poor performance.

To see how close to these benchmarks we should expect our evaluators to lie, Panels A and B of Appendix Figure E.2 plot the implied posteriors for male workers and female workers, respectively, *given evaluators' actual average confidence beliefs* from the *Baseline* treatment of the *Evaluator Study*. As such, these are the posterior beliefs that our evaluators would hold, given their beliefs, if they were Bayesian. As Appendix Figure E.2 makes evident, evaluators' over- and underconfidence beliefs are such that their implied Bayesian posteriors are almost exactly equal to their prior beliefs; that is, in our data, evaluators' confidence beliefs imply that they believe the signal to be almost entirely uninformative.

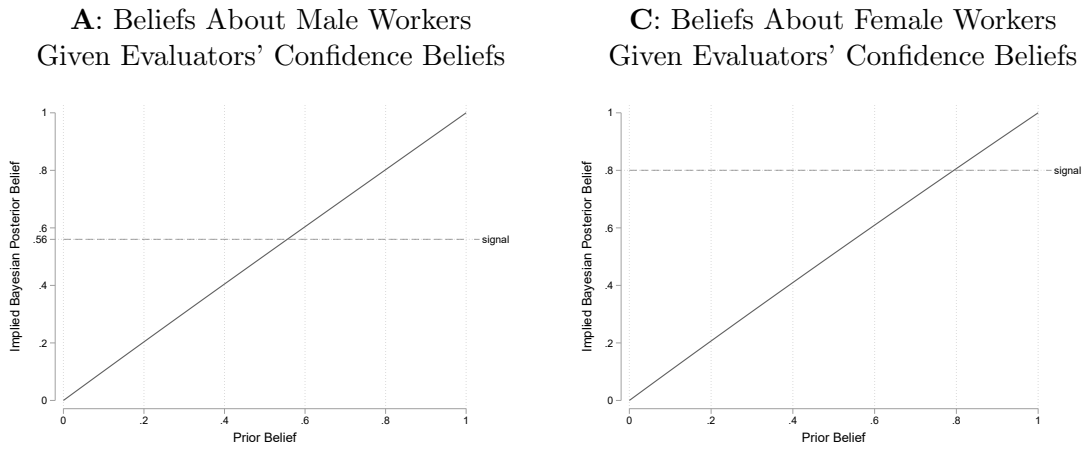
This is particularly striking in the context of our experiment. It implies that evaluators believe the signal to be as good as noise and therefore should discard it, but instead they incorporate it too much into their posterior beliefs. As a result, the gender gap in believed performance emerges from almost entirely uninformative signals.

One might worry that these implied beliefs instead result from confusion in the elicitation of the overconfidence and underconfidence beliefs, causing evaluators to naively answer 50%. First, even if this were to be the case, our main results are robust to this type of noise. Even without knowing the implied Bayesian posteriors, we can still say that evaluators are failing to account for

the gender gap in confidence since we find no difference between our main study and our *Unknown Gender* conditions. Second, even without the Bayesian posterior benchmark, it is still the case that evaluators fail to account for the gender gap relative to the true gap. Third, using another (unincentivized) elicitation, we still see that individuals who expect the gender gap in confidence do not account for it. Specifically, in our follow-up survey, we ask evaluators if they believe women to be less confident than men, and our results persist among the group of individuals who agree with this; see Section 5.1. Similarly, in our follow-up survey, we ask evaluators if they think that they accounted for the gender gap in confidence when making their predictions, and our results persist among the group of individuals who believe they did; see Section 5.2.

Finally, we note that two features of our confidence belief data indicate that evaluators did understand the confidence elicitation. First, less than 15% of evaluators report a belief of 50% and the distribution of beliefs is quite disperse (see Appendix Figure B.2 for histograms), so it is not the case that most evaluators respond with the heuristic of reporting 50%. Second, we find that confidence beliefs indeed indicate—as one may expect—that evaluators think male workers are relatively more overconfident than female workers and that female workers are relatively more underconfident than male workers.

Figure E.2: Implied Bayesian Posterior Beliefs as a Function of Evaluators' Confidence Beliefs



Graphs show the implied Bayesian posterior, across priors, given evaluators' beliefs about the likelihood that workers were over- and underconfident in the *Baseline* treatment of the *Evaluator Study*. Evaluators believed there to be a 39.86% chance that female workers were overconfident and a 48.11% chance that male workers were overconfident. They also believed there to be a 55.68% chance that female workers were underconfident and a 45.61% chance that male workers were underconfident. Bayesian updating is done separately for male workers and female workers based on the actual signal given to evaluators. When updating about male workers, evaluators are told that 56% of male workers believed that they had a poor performance. When updating about female workers, evaluators are told that 80% of female workers believed that they had a poor performance.